Testing Times: Challenges in Disentangling Admixture Histories in Recent and Complex Demographies

Matthew P. Williams[1*], Pavel Flegontov[2,3], Robert Maier[3], Christian D. Huber[1*]

1: Pennsylvania State University, Department of Biology, University Park, PA 16802, USA
2: Department of Biology and Ecology, Faculty of Science, University of Ostrava, Ostrava, Czechia
3: Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA

* Correspondence to Matthew. P. Williams: mkw5910@psu.edu, and Christian D. Huber: cdh5313@psu.edu

Abstract

Paleogenomics has expanded our knowledge of human evolutionary history. Since the 2020s, the study of ancient DNA has increased its focus on reconstructing the recent past. However, the accuracy of paleogenomic methods in answering questions of historical and archaeological importance amidst the increased demographic complexity and decreased genetic differentiation within the historical period remains an open question. We used two simulation approaches to evaluate the limitations and behavior of commonly used methods, qpAdm and the f_3-statistic, on admixture inference. The first is based on branch-length data simulated from four simple demographic models of varying complexities and configurations. The second, an analysis of Eurasian history composed of 59 populations using whole-genome data modified with ancient DNA conditions such as SNP ascertainment, data missingness, and pseudo-haploidization. We show that under conditions resembling historical populations, qpAdm can identify a small candidate set of true sources and populations closely related to them. However, in typical ancient DNA conditions, qpAdm is unable to further distinguish between them, limiting its utility for resolving fine-scaled hypotheses. Notably, we find that complex gene-flow histories generally lead to improvements in the performance of qpAdm and observe no bias in the estimation of admixture weights. We offer a heuristic for admixture inference that incorporates admixture weight estimate and *P*-values of qpAdm models, and f_3-statistics to enhance the power to distinguish between multiple plausible candidates. Finally, we highlight the future potential of qpAdm through whole-genome branch-length f_2-statistics, demonstrating the improved demographic inference that could be achieved with advancements in *f*-statistic estimations.

Keywords: aDNA, archaeogenetics, paleogenomics, qpAdm, *f*-statistics, admixture

Introduction

Beginning over a decade ago, the genome sequencing and analysis of ancient specimens, so-called ancient DNA (aDNA), spawned the field of paleogenomics and has provided novel insights into our understanding of population demographic history for a diversity of organisms and contexts (Brunson and Reich 2019; Spyrou *et al.* 2019; De Schepper *et al.* 2019; Arning and Wilson 2020; Mitchell and Rawlence 2021; Wibowo *et al.* 2021). No species has gained deeper insights from the aDNA revolution than humans, as it has significantly unraveled our complex evolutionary and migratory histories (Haber *et al.* 2016; Slatkin and Racimo 2016; Fu *et al.* 2016; Llamas *et al.* 2017; Williams and Teixeira 2020; Liu *et al.* 2021; Ávila-Arcos *et al.* 2023). Much of the research in human paleogenomics during the early 2010s was focused on reconstructing human prehistory (dating back more than 5k years before the present (YBP)) (Figure 1A). It was during these years that many of the statistical methods and software that have since become the foundation of aDNA studies were developed and have been pivotal in defining our understanding of human prehistory. These methods range from model-free exploratory approaches such as the smartpca implementation of principal component analysis (PCA) (Patterson *et al.* 2006; Reich *et al.* 2008; McVean 2009), and the ADMIXTURE software (Alexander *et al.* 2009), to statistical tests of admixture such as f_3- and f_4-statistics (Reich *et al.* 2009; Patterson *et al.* 2012), and the related D-statistics (Green *et al.* 2010; Durand *et al.* 2011), which leverage deviations from expected allele sharing patterns to reject simple trees and suggest more complex relationships. In addition, various downstream software has been developed to elucidate more complex relationships among numerous groups, with many utilizing *f*-statistics. Examples include qpAdm, which models a target population as a mixture of several proxy ancestry sources (Haak et al. 2015; Harney et al. 2021); qpWave, analyzing the number of gene flow events between population sets (Reich *et al.* 2012); and qpGraph, MixMapper, TreeMix, AdmixtureBayes, and findGraphs, all creating representations of admixture histories as directed acyclic graphs (Patterson *et al.* 2012; Pickrell and Pritchard 2012; Lipson *et al.* 2013, 2014; Nielsen *et al.* 2023; Maier *et al.* 2023). To a large degree, the reliance on these methods has been because of their use of allele frequencies which is suitable for pseudo-haploid aDNA whereby calling diploid genotypes is often infeasible due to its highly degraded characteristics.

Since the 2020s there has been a shift in aDNA research to studying the more recent past (Figure 1A). As a result, aDNA is increasingly used to address questions of archaeological and historical relevance. This research field was named archaeogenetics by British archaeologist Colin Renfrew (Boyle and Renfrew 2000). The historical period, particularly in Southwest Asia, is broadly demarcated to begin somewhere around the early-mid-3rd millennium BCE (Bartash 2020) and is characterized by the invention of writing, and intermittent periods of intensified inter-regional trade, diplomacy, and human mobility (Kristiansen 2016). From this body of research, hypotheses about gene flows between ancient settlements amenable to aDNA can involve groups separated by very short periods and thought to have descended from a complex web of migration and population structure (Haak *et al.* 2015; Lazaridis *et al.* 2016, 2017, 2022a; b; Haber *et al.* 2017, 2020; de Barros Damgaard *et al.* 2018; Harney *et al.* 2018; Wang *et al.* 2019; Narasimhan *et al.* 2019; Antonio *et al.* 2019;

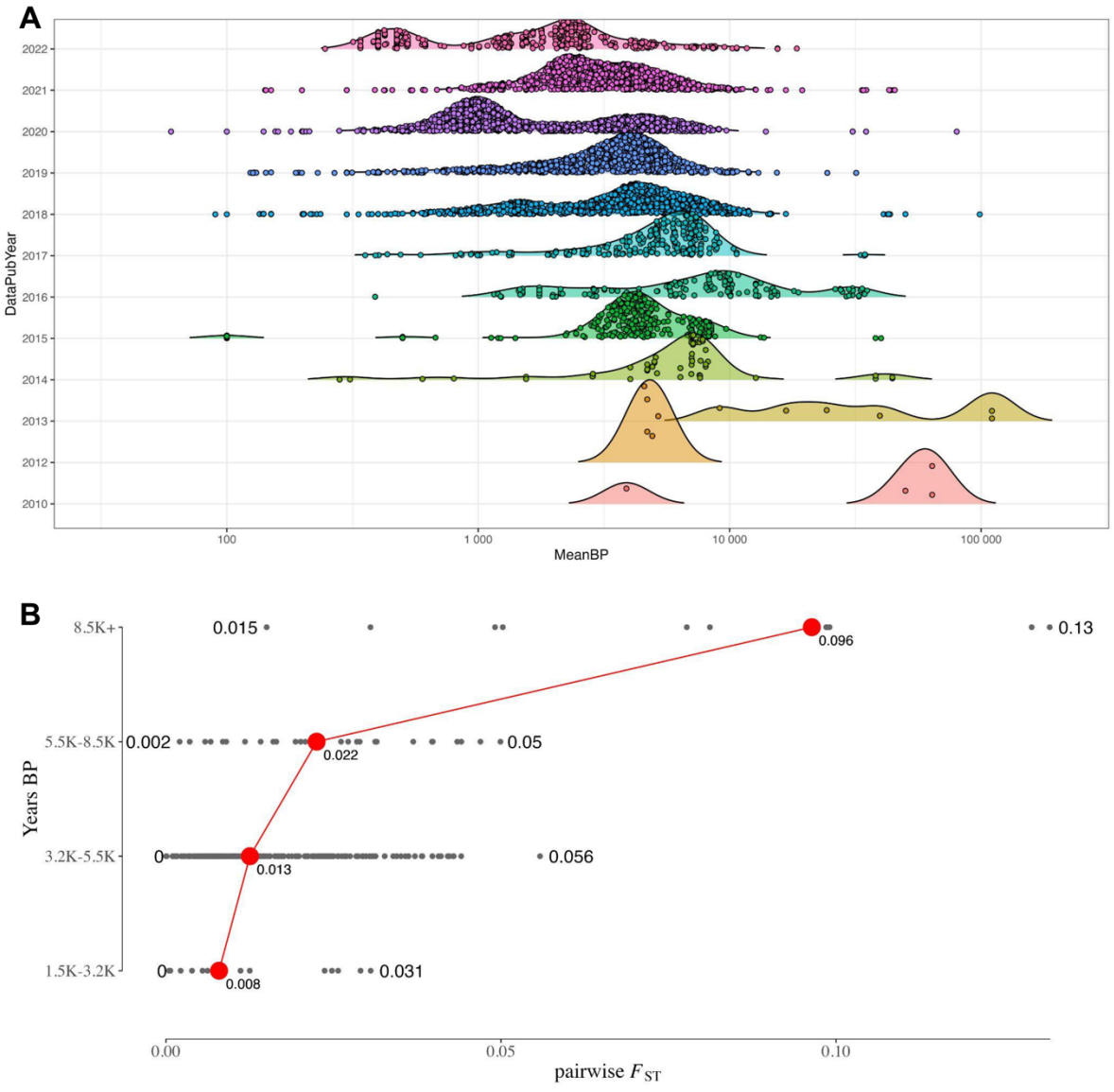
Fernandes *et al.* 2020; Agranat-Tamir *et al.* 2020; Skourtanioti *et al.* 2020, 2023; Clemente *et al.* 2021; Koptekin *et al.* 2023; Schmid and Schiffels 2023; Moots *et al.* 2023). These can range from questions regarding the degree of population continuity between periods of cultural change or settlement hiatus in the archaeological record to determining if cultural links between regions are indicative of inter-regional migration, and assessing if historical records of mass migrations and forced relocations result in observable signals of increased inter-regional gene flow. A common thread underlying these questions is, for a population of interest, to what extent can aDNA accurately reconstruct their genetic history, and importantly, reject false models of ancestry composed of closely related candidate populations? Moreover, what limits and possible biases emerge with the increase in demographic complexity amongst candidate source populations, a reduction in the number of generations separating aDNA samples and their ancestral admixture events, and an overall decrease in genetic differentiation indicative of the historical period? While the theoretical behavior of *f*- and D-statistics has been extensively tested (Patterson *et al.* 2012; Martin *et al.* 2015; Peter 2016, 2022; Harris and DeGiorgio 2017; Zheng and Janke 2018; Soraggi and Wiuf 2019; Tricou *et al.* 2022), and the performance of the commonly used software qpAdm thoroughly assessed under simple demographic models with both pulse-like and continuous migration (Ning *et al.* 2020; Harney *et al.* 2021), their behavior under varying degrees of population differentiation and complex demographic history expected of populations within the historical period remains underexplored.

In this study, we conducted a simulation-based evaluation of two widely used methods for reconstructing admixture histories - the "admixture" f_3-statistic and the qpAdm software (Figure 2). Our goal was to understand their effectiveness and limitations, particularly in complex scenarios that arise during the reconstruction of historical population dynamics. We started by simulating two chromosomes of combined length ~ 491 Mbp under four simplistic and qualitatively different admixture graphs, aiming to explore a broad range of model parameters leading to widely varying degrees of genetic differentiation. Subsequently, we expanded our evaluation to include a complex demography representative of a model of Eurasian human history emerging from a series of recent publications, which comprised 59 populations and 41 pulse admixture events. We simulated 50 whole-genome (L ~ 2875 Mbp) replicates and processed the simulated data to mimic typical aDNA conditions, including a Human-Origins-like SNP ascertainment scheme, empirical data missingness distributions, and pseudo-haploidization.

Importantly for the study of the historical period, our findings illustrate that qpAdm converges on a small subset of plausible models for an admixed target group consisting of the true sources and closely related populations by the time the F_{ST} levels reach those observed in Bronze and Iron Age Southwest Asian populations. However, under these divergence levels and conditions typical of aDNA, we observe qpAdm has limited ability to definitively answer fine-scaled questions relevant for archaeologists and historians due to lack of power to reject all non-optimal ancestry sources minimally differentiated from true ones. Moreover, for historical populations with complex gene-flow histories, we show that whilst admixture to source populations generally improves the performance of qpAdm, the phylogenetic origin of this admixture in ancestral source groups differentially

impacts qpAdm accuracy and performance. We show that the number of generations post admixture has no impact on qpAdm performance or accuracy of admixture proportion ("admixture weight") estimates. However, we observe when selecting sub-optimal ancestry sources that the admixture weights are biased in favor of the population that is most similar to the true source. We assessed several model plausibility criteria commonly used in the aDNA literature and show that each criterion impacts the performance and accuracy of qpAdm differently under various demographic conditions. Additionally, we highlight problems that users should be aware of when applying additional plausibility criteria for qpAdm models, such as negative admixture f_3-statistics or the rejection of all simpler qpAdm models, as they can lead to an increase in type II errors. Finally, we offer an interpretative heuristic guide that can enhance the power to distinguish between multiple plausible qpAdm models, thereby contributing to more robust and reliable archaeogenetic analyses.

Figure 1



Dates of published aDNA samples. (A) A per-publication-year transect of the density of the (log10) age of published ancient genomes. The publication dates and number of samples were taken from the Allen Ancient DNA Resource (AADR) v.52.2. (B) The temporal transect of population differentiation levels in Southwest Asia. The average dates for each sample in years BP were taken from the AADR v.52.2. For the plot in panel B, they were grouped into four epochs, with 3.2k years BP approximating the start of the Iron Age, 5.5k years BP approximating the start of the Bronze Age, 8.5k years BP approximating the start of the Neolithic period, and older years representing the Paleolithic period. The F_{ST} values were calculated using the Eigensoft v8.0.0 smartpca software.

Methods and Results

Starting simple: Insights into the behaviors of qpAdm and f_3-statistic from simplistic demographic models

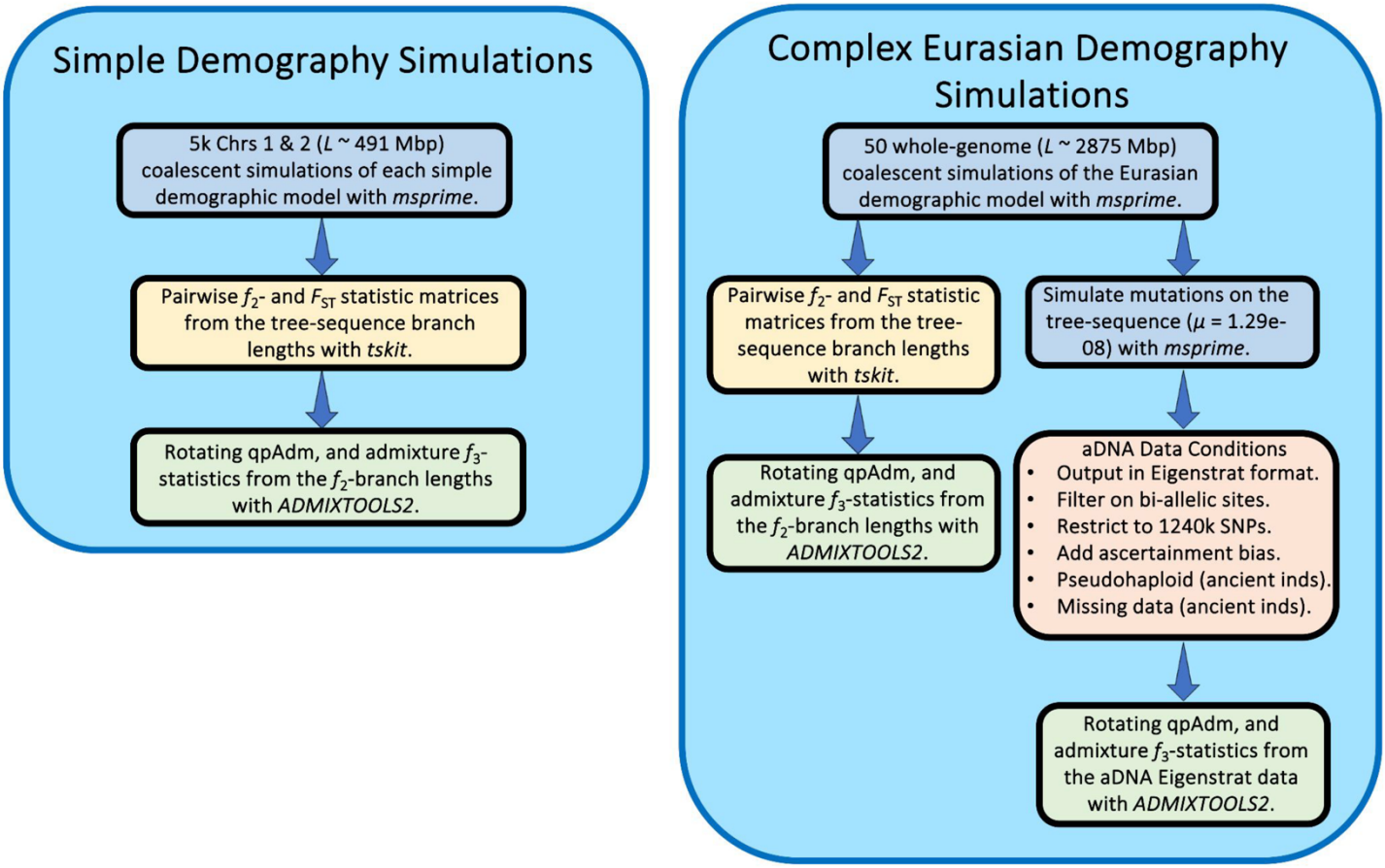
To obtain a baseline understanding of how specific demographic models and parameters impact downstream population genetic inference with qpAdm and the f_3-statistic, we formed simple bifurcating trees with varying scales of population divergence and augmented them with one to three gene flows in qualitatively different configurations (Figure 3A-D). For the simplest bifurcating demographic model with one admixture event (hereafter Model 1; Figure 3A), we randomly sampled values of five split-time parameters (T_1, T_2, T_3, T_4, and T_{admix}) from uniform distributions generated by the following framework:

- The oldest variable split-time (T_1) was selected first from a window between four generations in the past and the fixed T_0 split-time (6896 generations).
- The T_2 split-time parameter was sampled between three generations in the past and the sampled T_1 split-time.
- We selected the T_3 and T_4 split-time parameters from a window between two generations in the past and the T_2 split-time parameter.
- The T_{admix} (admixture date) parameter was selected from a window between a single generation in the past and the minimum of the T_3 and T_4 split-time parameters.
- We randomly sampled the admixture weight parameter (α), which forms the Target population as a mixture of the Source-1 (proportion α) and Source-2 (proportion 1 - α) populations, from a uniform distribution between zero and one (the distributions of simulated parameter values and scatter-plot matrices of simulation parameter correlations can be found in Supplementary Figure SI Figure S1A-B).

To assess the impact on admixture inference of more complex admixture history in one of proxy ancestry sources, we configured three additional demographic Models, each building upon the structure of Model 1 as follows:

- Model 2 includes a gene flow from an outgroup (R3 branch) into the source (S1).
- Model 3 includes admixture into the source (S1) from an internal branch ancestral to both the S2 and R2 populations (iS2R2).
- Model 4 combines the admixture events from Models 2 and 3, with no constraint on their order.

Figure 2



Simulation and analysis workflow in our study.

Data generation

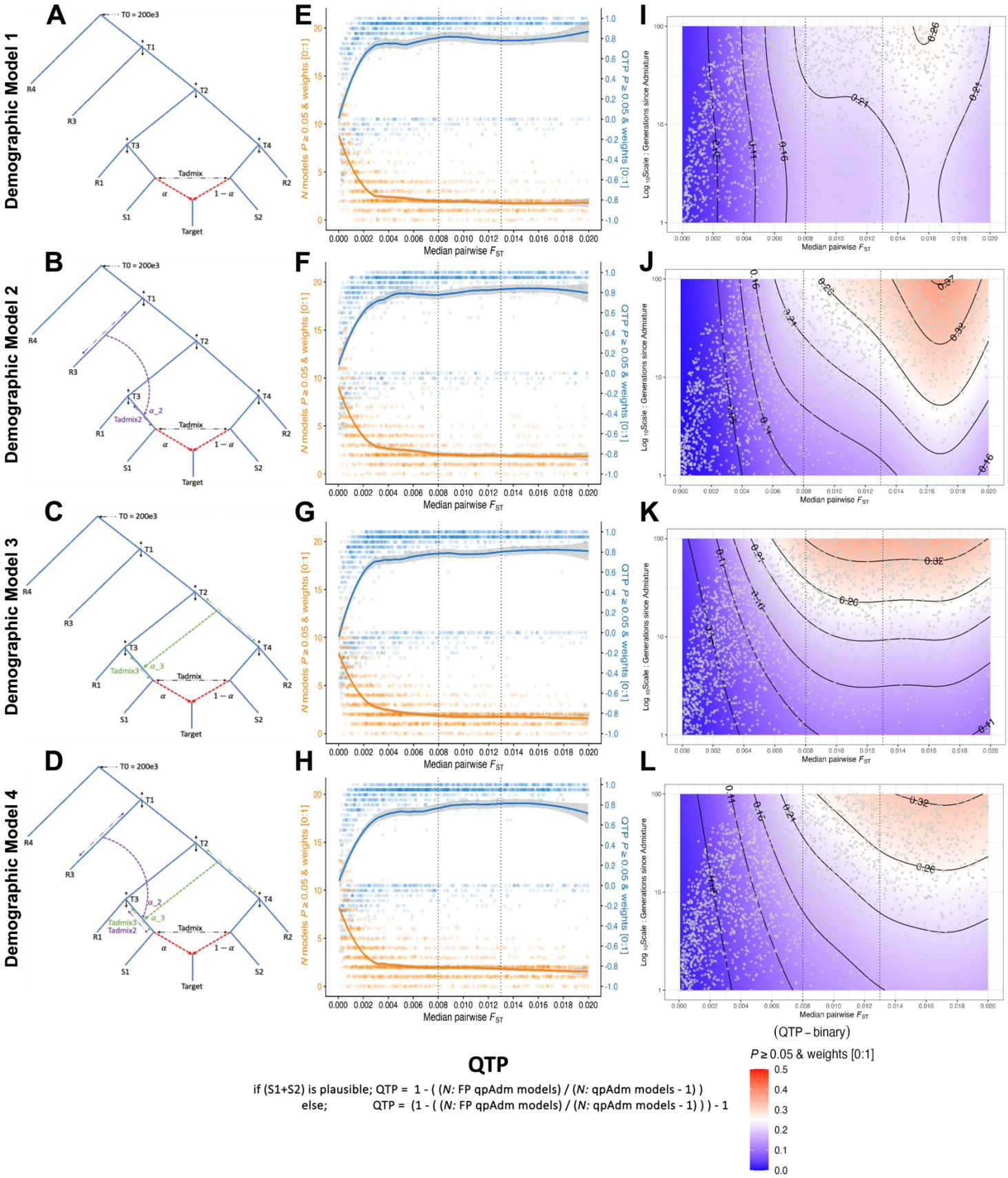
For each of the four simple demographic Models (Figure 3A-D), we used msprime v.1.2.0 (Kelleher *et al.* 2016; Baumdicker *et al.* 2021) to simulate 5000 iterations of succinct tree sequences without mutations with each iteration sampling demographic parameters from the schema outlined above. For the first 100 generations into the past we simulated under the Discrete Time Wright-Fisher model (DTWF) (Nelson *et al.* 2020), and then under the Standard (Hudson) coalescent model until the most recent common ancestor (MRCA). We used sequence lengths and recombination rates approximating human chromosomes one (L = ~2.49×10^8 bp, and r =

~1.15×10^{-8} per bp per generation) and two (L = 2.42×10^{8}, and r = 1.10×10^{-8}) (Adrion *et al.* 2020; Elise Lauterbur *et al.* 2022), and separated each chromosome with a log(2) recombination rate following guidelines in the msprime manual (https://tskit.dev/msprime/docs/stable/ancestry.html#multiple-chromosomes). For each demographic model, we fixed an upper bound split time of 200,000 years, and a generation time of 29 years, and for all populations, an effective size (N_e) of 10,00 and a sample size of 20 diploid individuals taken at the leaves.

We generated f_2 - and F_{ST} - statistic matrices directly from the tree sequences through tskit v.0.5.2 with parameters “Mode=branch”, and “span_normalise=True”, using 5 Mbp windows. The resulting f_2 - statistics matrix was used for qpAdm analyses with parameters “full_results=TRUE”, and “fudge_twice=TRUE”, and for calculating admixture f_3-statistics in the ADMIXTOOLS2 software (Maier *et al.* 2023). For the qpAdm rotating protocol following Harney et al. (2021), we included S1, S2, R1, R2, R3, and R4 as alternatively sources and “outgroups” (“right” populations), resulting in six single-source, and 15 two-source models. We computed admixture f_3-statistics on pairwise combinations of the S1, S2, R1, R2, R3, and R4 populations, resulting in an f_3-statistic test for each of the 15 two-source qpAdm models. The simple simulations resulted in genetic diversity estimates that cover ranges described for all present and past populations of anatomically modern humans, with the median pairwise F_{ST} between all Source and Right populations spanning from ~0.00012 to ~0.15.

Throughout our analysis of the simple demographic Models, we refer to the pairing of the S1+S2 Source populations as the “true” model representing the ancestry of the Target population, and we refer to all other population combinations as “false” models. In evaluating the qpAdm results, unless otherwise stated, we consider plausible models to have a P-value ≥ 0.05 and admixture weights between zero and one ([0:1]). In addition, we configured a summary metric, “qpAdm test performance” (QTP), that conveys the precision of rotating qpAdm analyses per simulation iteration taking into account all single and two-source qpAdm models (Figure 3E-H). The range of QTP is between “+1” and “-1” where the most optimal outcome, “+1”, corresponds to the condition where all false models are rejected (single and two-source) and the true model is plausible. The worst outcome, “-1”, occurs when the true model is rejected, and all false models are considered plausible. As such, all rotating qpAdm analyses that reject the true model result in a negative QTP, and analyses that have the true model amongst the plausible qpAdm models have positive QTP values. The outcomes where all models are rejected, or all models are plausible are scored as “0”. Values between “+1” and “0” occur when both the true and false modes are plausible in the same simulation, with each additional plausible false model (single and two-source) decreasing the QTP value. Likewise, values between “0” and “-1” occur when the true model is rejected, and some (but not all) false models are plausible. We also evaluated the binary QTP outcome (Figure 3I-L), whereby qpAdm either performs most optimally (i.e. rejects all false models and estimates the true model as plausible) or does not (i.e. at least one wrong model is considered plausible or the true model is rejected).

Figure 3



Simple demographic models and qpAdm test performance (QTP). (A-D) Topological structures of the four simple demographic models. (E-H) QTP and number of plausible qpAdm models across the range of median pairwise F_{ST} values calculated on the S1, S2, R1, R2, and R3 populations. For each simulation iteration we represent the counts of the number of plausible single and two-source qpAdm models (21 is the maximum possible) with orange dots and the locally estimated scatterplot smoothing (loess) computed in R and shown with the orange line. We show the QTP value for each simulation iteration with blue dots and the loess smoothing with the blue line. (I-L) Logistic GAM probability for the QTP-binary response variable with admixture date (T_{admix}) and median pairwise F_{ST} as predictor variables. The gray dots are unique combinations of simulation parameters placed in the space of predictor variables. Vertical dotted lines in plots E-L show the median pairwise F_{ST} values at the approximate Iron (0.008), and Bronze Age (0.013) periods.

The Limits of Population Differentiation for qpAdm Admixture Model Inference

Due to extensive admixture between ancient southwest Eurasian groups beginning around the 6th millennium BCE, populations from historical periods exhibit, on average, lower genetic differentiation than their predecessors (Figure 1B). Therefore, evaluating archaeogenetic hypotheses regarding historical migrations necessitates the ability to disentangle the admixture histories of minimally differentiated ancient groups separated by very short periods of genetic drift. To address this, we used demographic Model 1 (Figure 3A) to directly evaluate the impact and limits of population differentiation on the performance of rotating qpAdm. For all downstream demographic inference analyses, we constrained the simulated parameter space to values that approximate conditions observed amongst historical period groups such as a low median pairwise F_{ST} between 0 and 0.02 computed on the S1, S2, R1, R2, and R3 populations, and ≤ 100 generations since the admixture event forming the Target population. Unless otherwise stated, we use this parameter range for all results described below.

Genetic differentiation and qpAdm performance

A requirement of qpAdm is that at least one right-group population is differentially related to populations in the left set (Haak *et al.* 2015; Harney *et al.* 2021) as the power of qpAdm is largely due to the right-group populations' ability to distinguish between putative ancestry sources (Harney *et al.* 2021). Consistent with this principle, we observe a general trend of increasing qpAdm performance (QTP) with larger median pairwise F_{ST} values (Figure 3E). As these values approach 0.01, equivalent to that observed amongst Southwest Asian Bronze Age and older groups, we notice QTP to asymptote around 0.8 and convergence on an average of two plausible qpAdm models per simulation iteration (Figure 3E). However, as the median pairwise F_{ST} drops to values observed at the lower ends of human population differentiation (~0.003 - 0.004) we observe a sharp decline in the average QTP driven by increases in both the number of plausible false qpAdm models and rejections of the true qpAdm model (S1+S2) (Figure 3E).

To analyze the distribution of plausible qpAdm models driving the QTP variation at different levels of genetic differentiation, we formed median pairwise F_{ST} bins roughly corresponding to values separating historical epochs. The smallest range, F_{ST} between 0 and 0.008, corresponds to the diversity estimated from samples dating between 1.5k to 3.2k years ago (Figure 1B) with the upper range broadly demarcating the Iron Age from the Bronze Age in Southwest Asia. The middle range, F_{ST} between 0.008 and 0.013, corresponds to the diversity estimated from samples dating between 3.2k and 5.5k years ago and encompasses the Bronze Age population diversity (Figure 1B). The upper range, F_{ST} between 0.013 and 0.02, estimated from samples dating between 5.5k and 8.5k years ago, represents the diversity present amongst populations ancestral to those of the historical period (Figure 1B). Consistent with the QTP distribution described above, the smallest F_{ST} bin contains the highest number of false plausible qpAdm models including single-source models for the target population (Figure 4A). The degree of population divergence also impacts the plausibility of the true model with larger F_{ST} bins increasing both the frequency of plausible true models (0.705, 0.859, and 0.842 for the three F_{ST} bins, respectively) and the proportion of true models out of all plausible qpAdm models (22.5%, 44.8%, and 48.7%, for the three F_{ST} bins, respectively). Notably, the increased rejection of the true model in the lowest F_{ST} bin is largely due to inaccurate estimations of the admixture weights. Approximately 50% of the true model replicates with *P*-values ≥ 0.05 are rejected due to admixture proportions outside the [0:1] range (SI Figure S2).

For larger F_{ST} bins, the predominant rejection of the true model shifts to statistical significance, with the majority of true models rejected with *P*-values between 0.01 and 0.05 (SI Figure S2).

With the recent shift of aDNA research towards reconstructing admixture histories within sub-continental regions (Ávila-Arcos et al. 2023), understanding the limits of rejecting false sources recently split from the true ancestral source is becoming increasingly pertinent. To investigate this, we explored the limits of differentiating between the sister clades of R1 and S1, and by symmetry S2 and R2, (Figure 3A) as false and true sources in qpAdm models. As expected, qpAdm has the greatest difficulty rejecting models that combine one of the (false-source) cladal populations with one of the true sources, as combinations of S1+R2 and S2+R1 account for more than 25% of all plausible qpAdm models across all F_{ST} bins (Figure 4A). As anticipated given the topological symmetry of Model 1, the two false qpAdm models are plausible at almost equal frequency. However, we less frequently observe that both false models are plausible within the same simulation (SI Figure S3), consistent with the convergence towards an average of two plausible qpAdm models described above (Figure 3E).

To assess the relationship between genetic differentiation within putative source clades and performance of rotating qpAdm, we analyzed the joint distribution of S1-R1 and S2-R2 F_{ST} values for all false qpAdm models that included one of the R1 or R2 populations. As expected, we observe on average larger F_{ST} values between the S1-R1 and S2-R2 populations for rejected false models (mean = 0.004, and median = 0.002) than plausible false models (mean = 0.002, and median = 0.001) which resulted in statistically significant differences between their respective F_{ST} distributions (Mann-Whitney U *P*-value < 0.001) (SI Figure S4). Thus, our simulations

suggest that under the simple topological structure of Model 1, rotating qpAdm has the power to differentiate between closely related cladal populations, albeit with more difficulty distinguishing between putative sources separated on the order of F_{ST} < ~ 0.002.

Table 1

qpAdm Average Performance: Historical Simulation Parameters

Plausibility Criteria	Model 1				Model 2				Model 3				Model 4			
	FP	FDR	QTP	QTP-binary	FP	FDR	QTP	QTP-binary	FP	FDR	QTP	QTP-binary	FP	FDR	QTP	QTP-binary
P-value 0.01	0.318	0.844	0.644	0.000	0.301	0.806	0.658	0.003	0.330	0.842	0.639	0.000	0.321	0.839	0.647	0.001
P-value 0.05	0.275	0.841	0.598	0.000	0.255	0.795	0.627	0.009	0.282	0.830	0.624	0.004	0.274	0.829	0.616	0.004
P-value 0.01 + weights [0:1]	0.119	0.589	0.761	0.121	0.128	0.592	0.762	0.129	0.118	0.594	0.718	0.125	0.111	0.583	0.724	0.137
P-value 0.05 + weights [0:1]	0.098	0.574	0.699	0.148	0.107	0.574	0.712	0.157	0.097	0.566	0.683	0.166	0.092	0.566	0.676	0.155
P-value 0.01 + weights [0:1] ± 2s.e	0.074	0.624	0.608	0.117	0.081	0.610	0.640	0.131	0.067	0.644	0.521	0.116	0.065	0.628	0.537	0.127
P-value 0.05 + weights [0:1] ± 2s.e	0.060	0.610	0.554	0.143	0.066	0.592	0.597	0.159	0.054	0.614	0.495	0.152	0.052	0.615	0.495	0.142
All single-source models rejected																
P-value 0.01 + weights [0:1]	0.070	0.631			0.074	0.631			0.065	0.650			0.064	0.630		
P-value 0.05 + weights [0:1]	0.060	0.609			0.065	0.605			0.056	0.611			0.055	0.606		
P-value 0.01 + weights [0:1] ± 2s.e	0.067	0.642			0.070	0.630			0.062	0.663			0.060	0.643		
P-value 0.05 + weights [0:1] ± 2s.e	0.057	0.622			0.060	0.606			0.051	0.625			0.050	0.623		
All single-source models rejected & significant f3-statistics																
P-value 0.01 + weights [0:1]	0.052	0.618			0.056	0.601			0.053	0.655			0.051	0.633		
P-value 0.05 + weights [0:1]	0.043	0.595			0.045	0.574			0.043	0.625			0.042	0.608		
P-value 0.01 + weights [0:1] ± 2s.e	0.052	0.625			0.055	0.604			0.052	0.664			0.051	0.643		
P-value 0.05 + weights [0:1] ± 2s.e	0.043	0.603			0.045	0.576			0.042	0.634			0.041	0.621		

Performance summaries of *qpAdm* rotation analysis for the four demographic models and different performance metrics. Each cell contains the average of each performance metric under different qpAdm plausibility criteria. The averages are over the parameter range of the historical period (admixture generations ≤ 100 and median pairwise F_{ST} > 0 and ≤ 0.02). The performance metrics are as follows: FP = false positive rate, FDR = false discovery rate, QTP = qpAdm test performance, and QTP-binary = qpAdm test performance provided that only the true model fits the data. Each performance metric is evaluated under a different model plausibility criteria for the four demographic Models. Their averages are printed in each cell with the color ranging from smaller (yellow) to medium (green), and larger (purple) values. For each plausibility criteria we highlight with a red box the demographic Model that performed the best for each performance metric. For example, at the plausibility criteria of *P*-value ≥ 0.01 and the FP metric, Model 2 has the smallest value and thus performed the best and a red box around its cell.

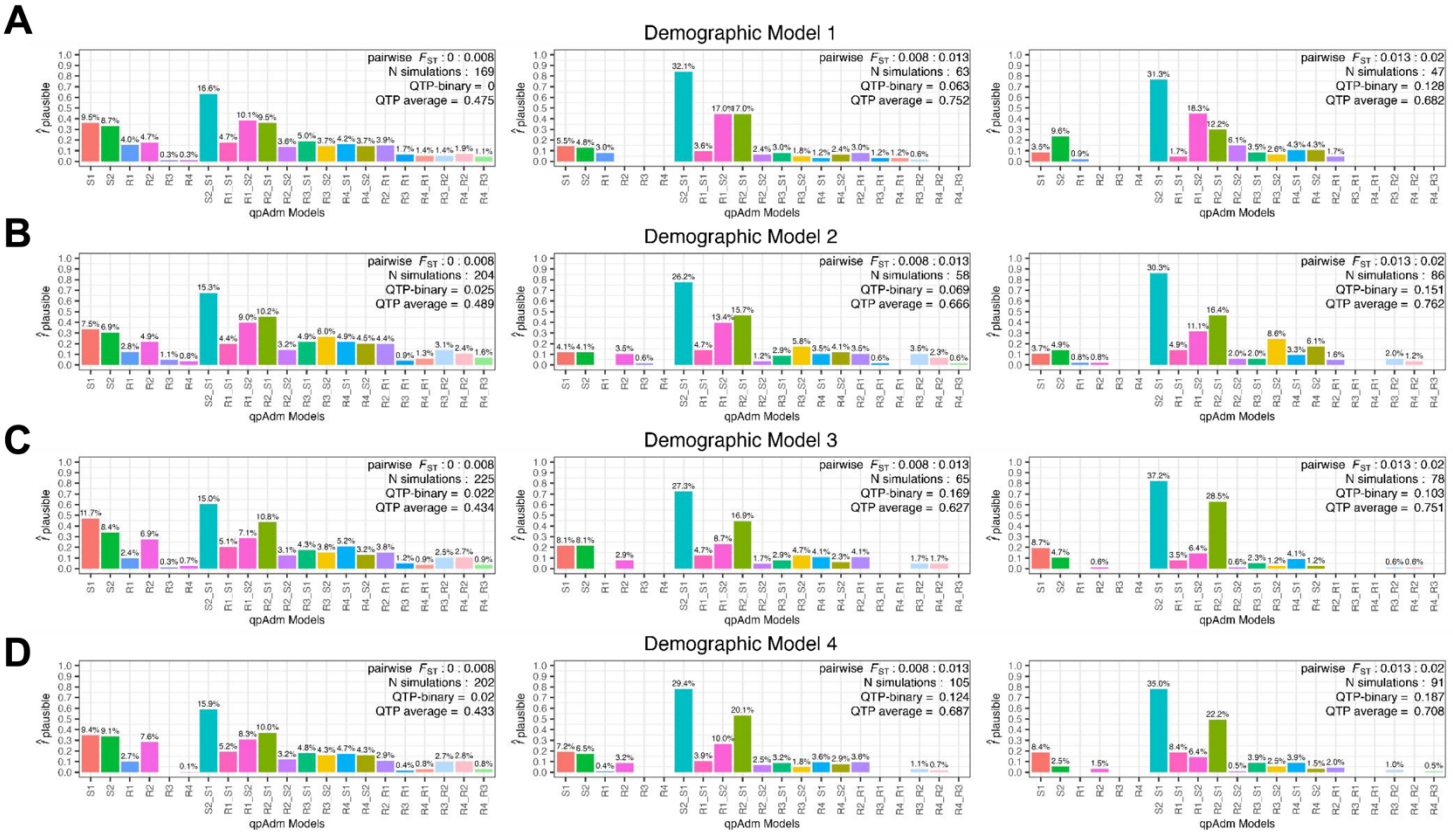
More Complex Admixture History of Sources Affects Demographic Inference

Often, complex ancestral relationships exist among putative historical source populations (e.g., Lazaridis et al. 2016), however, whether this is detrimental to the effectiveness of identifying admixture patterns through qpAdm remains unknown. To assess the impact of both the introduction, phylogenetic origin, and number of admixture events into the source population on admixture inference we performed 5,000 simulations on each of three demographic Models that introduce admixture to the source (S1) population (Figure 3), with all other simulation parameters remaining consistent with Model 1.

Importantly, the addition of admixture events into the S1 population does not lead to significant changes to the distribution of QTP across the F_{ST} range. Results for all demographic Models converge on a maximum average QTP of ~ 0.8 and an average of two plausible qpAdm models (Figure 3E-H). We do observe subtle differences in their average performance for metrics such as False Positive Rate (FPR) = FP / (FP+TN), False Discovery Rate (FDR) = FP / (FP+TP), QTP, and QTP-binary (Table 1). From each simulation iteration, we computed the qpAdm FPR for each demographic Model as follows: we counted the number of plausible false qpAdm models (false positives: FP) to obtain the FP qpAdm model count. To obtain the number of true negative (TN) qpAdm models, we counted the number of rejected false qpAdm models. For example, in Model 1 simulation iteration 1,998, we have an FPR of 0.8 that occurred because, of the 21 total single and two-source qpAdm models, we have 16 FP qpAdm models and four false qpAdm models were rejected (FP / (FP+TN) = 16 / 20). We computed the FDR in the same fashion. The observation of a plausible true qpAdm model represents the true positives count (TP), meaning an FDR of 1 occurs when only false qpAdm models are plausible and 0 when only the TP qpAdm model is observed and all false qpAdm models are rejected. We then averaged these metrics to generate a summary of the overall performance for each Model under historical period parameters. No single demographic Model consistently outperforms others across all performance metrics, indicating that different admixture scenarios have varying effects on qpAdm performance and accuracy. This is further supported by the observation that across multiple model plausibility criteria (discussed further below), the average QTP, QTP-binary, and FDR consistently favor demographic Model 2, while the FPR is most frequently lowest for Model 4 (Table 1). However, we note that the best-performing average qpAdm metric consistently falls within one of the more complex Models 2 to 4, suggesting that, on average, the introduction of admixture to the Source population increases qpAdm rotation performance even though it decreases overall population differentiation (both median and average F_{ST} is largest in Model 1).

Similarly to Model 1, all Models with complex admixture history of S1 exhibited the highest number of false-plausible qpAdm models in the lowest range of population divergence (Figure 4). However, while we observed very similar frequencies of plausible S1-R2 and S2-R1 false qpAdm models under demographic Model 1, the introduction of admixture to the S1 population introduced an asymmetry, with the S2-R1 qpAdm model being more frequently rejected than the S1-R2 model (Figure 4B-D). Interestingly, this asymmetry is most pronounced under Model 3, which involves admixture from the common ancestor of S2 and R2 (iS2R2) to the S1 branch, and the asymmetry further increases in larger F_{ST} bins (Figure 4C). Demographic Model 4 including two admixture events in S1, displayed a distribution of false plausible models across sources intermediate between Models 2 and 3 (including one admixture event in S1) suggesting the phylogenetic source of gene flow in S1 has a greater impact on the resulting plausible qpAdm models than the number of admixture events in S1 (Figure 4B-D). This has important implications for the empirical study of ancient populations whose sources are themselves admixed (see Discussion).

Figure 4



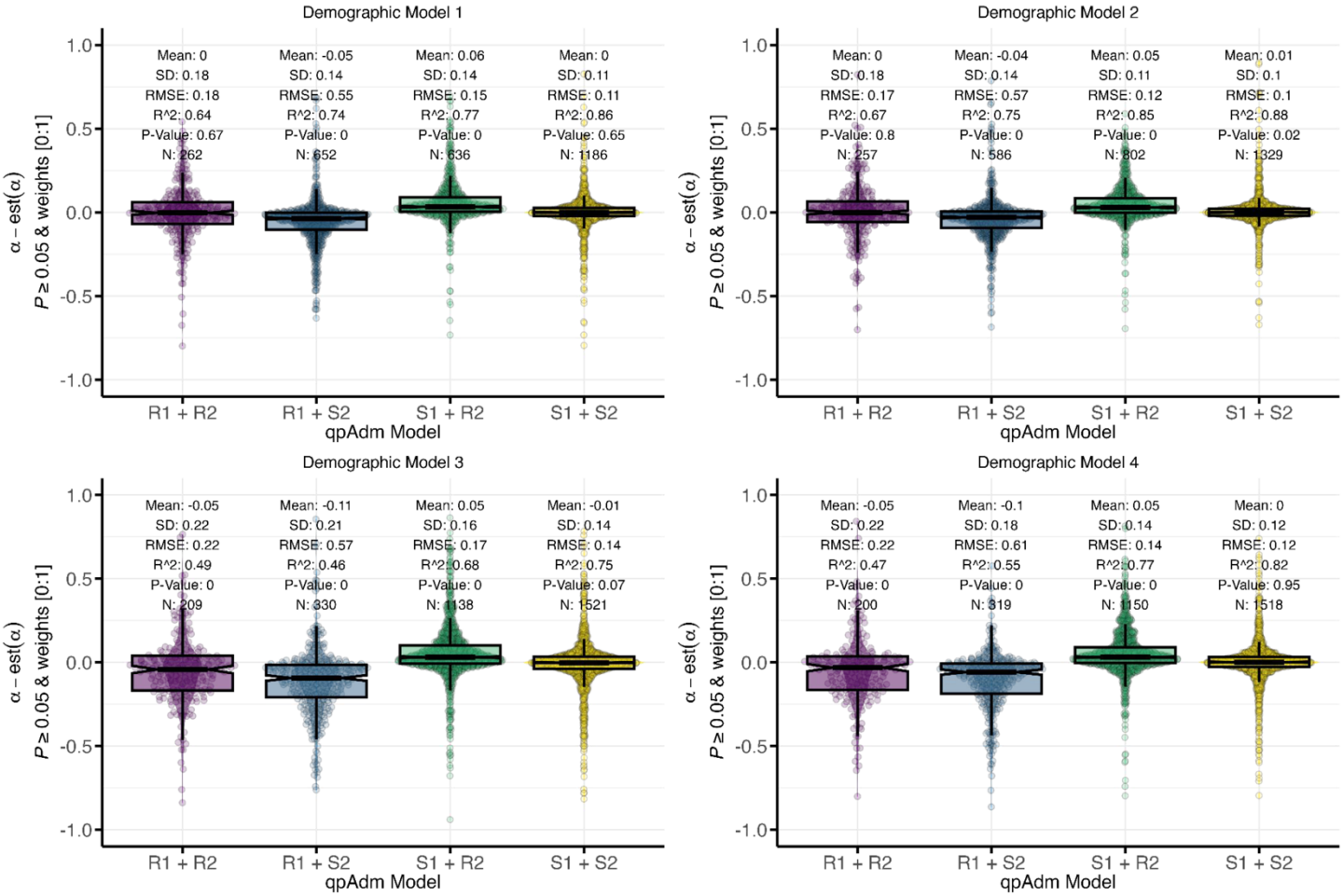
Distribution of plausible one-source and two-source qpAdm models across three population differentiation ranges (between the S1, S2, R1, R2, and R3 populations) and across the four simple demographic models. The number of generations since admixture is less than or equal to 100 in all cases. Each row represents one simulated demographic history and the columns are increasing ranges of population differentiation (F_{ST}) corresponding to the historical period demarcations indicated in Figure 1B. The values above each barplot represent the proportion of all plausible qpAdm models within the simulation iterations for each differentiation range. The y-axis shows the frequency of each model as plausible across the total number of simulations within each differentiation range. In the top right-corner of each barplot is shown the F_{ST} range, number of simulations within that range, and the average QTP and QTP-binary.

A further challenge in the use of aDNA in resolving hypotheses regarding migrations during historical periods is the increased likelihood of studying recent admixture events. Such scenarios may arise in the context of detecting shifts in genetic ancestry after episodes of human migration, where only a few generations separate the timing of admixture and the ancient human individuals sampled. Moreover, the effectiveness of qpAdm in addressing historical questions that necessitate the identification of a specific population or lineage responsible for admixture is inversely proportional to the number of plausible models it identifies. We assessed the performance of qpAdm under both of these challenges by modeling the interaction between generations since admixture and population divergence on the probability of exclusively identifying the true qpAdm model using a logistic generalized additive model (GAM) in the mgcv v.1.9.0 R package (Wood 2004) with QTP-binary as the response variable, automatic smooth terms for each of the predictor variables (median pairwise F_{ST} between the S1, S2, R1, R2, R3 populations; generations since admixture), and the model parameters were estimated using restricted maximum likelihood (REML). The model's output was in the form of log-odds, which we then

converted to probabilities. This conversion was done by first exponentiating the log-odds to get the odds ratio, and then dividing the odds ratio by one plus the odds ratio (i.e., Probability of QTP-binary = odds-ratio / (1 + odds-ratio)). We visualized these predicted probabilities on a grid that represents the space of the historical parameters (Figure 3I-L).

As expected, larger median pairwise F_{ST} values resulted in increased QTP-binary probability for all simple demographic Models (Figure 3I-L), with the more complex Models 2-4 performing better than Model 1 across historical F_{ST} ranges. Counterintuitively, the model with admixture from the internal branch ancestral to both the S2 and R2 populations (iS2R2) (Model 3) performed the best at F_{ST} values both below (median pairwise F_{ST} < 0.008) and within (median pairwise 0.008 < F_{ST} < 0.013) ranges approximating that of historical periods (Figure 3I-L). When median divergence levels reached those of populations older than the Bronze Age (median pairwise F_{ST} > 0.013) the model with a gene flow from an outgroup to a source branch (Model 2) outperformed the others, achieving the same QTP-binary probability values with fewer generations since admixture than Models 3 and 4 (Figure 3J). It also had the highest maximum QTP-binary probability of all Models, achieving this with generations since admixture greater than ~90 (Figure 3J). In the absence of admixture events in the history of S1 (Model 1), we observed no significant impact of generations since admixture on the QTP-binary probability (Chi-sq = 2.25 and *P*-value = 0.089). However, all three admixed-source Models, especially Model 3, show a weak but statistically significant effect of generations since admixture on QTP-binary, with the effect appearing more pronounced for larger F_{ST} values (approximate significance of T_{admix} predictor variable smooth term: Model 2 Chi sq = 9.94 and *P*-value = 0.0012; Model 3 Chi sq = 23.79 and *P*-value < 0.001; and Model 4 Chi sq = 10.17 and *P*-value = < 0.001.) (Figure 4I-L). The observed weak influence of generations post-admixture on the QTP-binary probability is likely a consequence of correlations between the T_{admix} and T_3/T_4 parameters (SI Figure S1B) rather than a decline in performance due to more recent admixture. In support of this idea is that both Models 3 and 4, which incorporate admixture from the iS2R2 branch that is delineated by the T_2 and T_4 split times (Figure 3C-D), have the strongest correlation between T_{admix} and T_4 of all demographic Models (SI Figure S1B). Conversely, under Model 2, T_{admix} has the strongest correlation with parameter T_3 (SI Figure S1B), which determines the divergence time between R1 and S1.

Figure 5



Deviations of estimated from simulated admixture proportions for the R1 and S1 sources in the *qpAdm* models S1+S2, R1+R2, R1+S2, and S1+R2. Median pairwise F_{ST} between the S1, S2, R1, R2, and R3 populations is between 0 and 0.02 and the number of generations since admixture is less than or equal to 100. Each panel shows results for one simple demographic model.

Accuracy of Admixture Weight Estimates

We also evaluated if the introduction of admixture to the ancestral source population (S1) would introduce bias or increase uncertainty in the admixture weight estimation of the Target for the true qpAdm model. Consistent with previous studies (Harney *et al.* 2021), in the absence of ancestral admixture, we observed a delta alpha (simulated minus estimated admixture weight) mean of zero under Model 1 and an R^2 of 0.86 demonstrating qpAdm can accurately estimate the simulated admixture weight without bias (one-sample T-test *P*-value = 0.65) (Figure 5). However, in the presence of admixture to the source population from an outgroup, we observe a subtle and weakly significant overestimation of the S1 contribution to the Target (Model 2, delta-alpha mean = 0.01, one-sample T-test *P*-value = 0.02), and an underestimation of almost equal magnitude, but not significant, when admixture in S1 is from the iS2R2 branch (Model 3, delta-alpha mean = -0.01, one sample T-test *P*-value = 0.07) (Figure 5). The symmetrical biases between Model 2 and Model 3 appear to cancel out under demographic Model 4, where we observe a delta-alpha mean of zero (one-sample T-test *P*-value = 0.95)

(Figure 5). All demographic Models exhibited similar levels of uncertainty in their admixture weight estimation (Figure 5). Whilst Model 3 performed the worst with the lowest R^2, largest delta-alpha standard deviation, and root-mean-squared error (Figure 5), the weight-estimate uncertainty is considerably smaller than expected under completely random sampling (the SD of the difference between two uniformly distributed and uncorrelated random variables is 0.408) further supporting the accuracy of qpAdm admixture estimates under these conditions.

In empirical aDNA studies, one will often include multiple closely related populations in the qpAdm candidate source list to determine which is the best representation of the Target ancestry. Therefore, we assessed how the selection of false sources and their phylogenetic relationship to the true source affected the bias and uncertainty in admixture weight estimates. Under the simplest model (Model 1), we observed that misspecified (false) models that combine the true Source populations with the sister clade of the other true Source (R1+S2 or S1+R2) resulted in an almost equal overestimation of the simulated admixture weight for the true Source (R1+S2 mean = -0.05, T-test *P*-value = < 0.001 whereas S1+R2 mean = 0.06, T-test *P*-value < 0.001) (Figure 5). However, when both sources are equally phylogenetically distant from the true admixing sources (R1+R2), we only observed an increase in the weight estimate uncertainty but no bias (SD = 0.18 and T-test *P*-value = 0.67) (Figure 5). We observed similar qualitative patterns in the admixed-source models (Models 2-4), with a bias in overestimating the contribution from the true source when paired with one of the false sources (S1+R2 and S2+R1 T-test *P*-values < 0.001) (Figure 5). Interestingly, the largest effects are observed under demographic Models 3 and 4 for the qpAdm model S2+R1, suggesting that admixture from the internal iS2R2 branch to the ancestral S1 branch increases the overestimation of the S2 contribution to the Target (Figure 5). Moreover, selecting the two symmetrical populations R1+R2 results in an overestimation of the R2 contribution under both Models 3 and 4 (T-test *P*-value < 0.001), but we observed no bias under Model 1 (T-test *P*-value = 0.67). Additionally, the Model with a gene flow from an outgroup to the ancestral S1 branch (Model 2) has no bias in admixture weight estimation (R1+R2 T-test *P*-value = 0.8), further supporting the impact of the admixture between ancestral source branches (iS2R2) on the qpAdm weight bias (Figure 5).

qpAdm Plausibility Criteria and Improving Model Inference Accuracy

A number of different qpAdm plausibility criteria are employed in empirical aDNA analysis such as *P*-value thresholds of 0.01 (e.g., Skoglund et al. 2017, Narasimhan et al. 2019, Lazaridis et al. 2022, Bergström et al. 2022, Koptekin et al. 2023, Skourtanioti et al. 2023) and 0.05 e.g., (Olalde *et al.* 2019; Sirak *et al.* 2021; Salazar *et al.* 2023), the use of two-standard error constraint (S.E.) on the admixture weights (Narasimhan *et al.* 2019), the requirement of the rejection of all single-source models, and favoring simpler models over more complex ones (Lazaridis *et al.* 2016, 2022a; Skoglund *et al.* 2017; Narasimhan *et al.* 2019; Salazar *et al.* 2023). Our objective was to determine how these plausibility criteria impact the performance (QTP and QTP binary) and accuracy (FPR and FDR) of qpAdm admixture model inference. Additionally, we aimed to assess whether the

accuracy (FPR and FDR) of qpAdm could be improved by conditioning on a significant admixture f_3-statistic for plausible two-source models, a method used to test if a target population is consistent with being formed from two putative sources (Patterson *et al.* 2012; Peter 2016, 2022).

We observed a substantial decrease in the average error rates (FPR and FDR), and an increase in average performance metrics (QTP, and QTP-binary) across all demographic Models when introducing the admixture weight [0:1] constraint to the plausibility criteria in qpAdm (Table 1). We note that the admixture weight [0:1] constraint is the most common additional constraint on qpAdm model plausibility in the archaeogenetic literature. However, adding the additional ± 2 S.E. weight constraint, while reducing the FPR for all demographic Models, also increased the FDR for both *P*-value thresholds (Table 1), highlighting the trade-off between rejecting the true model and failing to reject false models when assessing accuracy. Similarly, the ± 2 S.E weight constraint also decreased the average QTP results for both *P*-value thresholds across all Models and only Model 2 shows an increase in average QTP-binary (increases in QTP-binary for both *P*-values 0.01 and 0.05) (Table 1).

We evaluated the impact of requiring all single-source qpAdm models to be rejected on the FP and FDR error rates. We computed the FPR as follows: For each simulation iteration, if at least one false single-source qpAdm model was plausible, all two-source qpAdm models were rejected and we then computed the FPR as the FP /(FP + TN) following the guide above. Meaning, a two-source qpAdm model can only contribute to the FPR if all single-source qpAdm models are rejected in its simulation iteration. Recalling the above example, in the demographic Model 1 simulation iteration 1,998, we had an FPR of 0.8 that occurred because, of the 21 total single and two-source qpAdm models, we have 16 FP models, six of which are single-source models. However, because we conditioned on all single-source models to be rejected, we have six false positives (single-source models) and 14 true negatives (rejected two-source), giving this particular simulation iteration an FPR of 0.3. The FDR was computed following the same procedure, where all two-source qpAdm models are rejected if a single-source model is plausible in their simulation iteration. Importantly, we found that requiring all single-source models to be rejected increased the FDR for all demographic Models at both *P*-value thresholds (Table 1). Conversely, we find that the FPR is decreased with the rejection of all single-source models for all demographic Models across all plausibility criteria (Table 1).

In addition to rejecting all single-source qpAdm models, the further criterion of a significant admixture f_3-statistic for plausible two-source qpAdm models resulted in the lowest error rates (FPR and FDR) for all demographic Models (Table 1). The relationship between the power of the admixture f_3-statistic and demographic parameters was explored by (Peter 2016). They showed through mathematical formulae (see equation 1 below) and simple simulations similar to our Model 1, that the conditions of a negative f_3-statistic required a large number of generations between the split of the admixing sources (T_2) and the time of admixture (T_{admix}), a low probability of

lineages in the Target population coalescing before the admixture event (T_{admix}), and the admixture proportion (α) close to 50%. As such, for any pair of true-source populations to produce a negative f_3-statistic for a target, the demographic model from which they descend must conform to the equation (1) (EQ:1) below.

$$\text{negative } f_3\text{-statistic condition} = \left(\frac{1}{(1-c_x)}\frac{T_{admix}}{T_2} < 2\alpha(1-\alpha)\right) \quad (1)$$

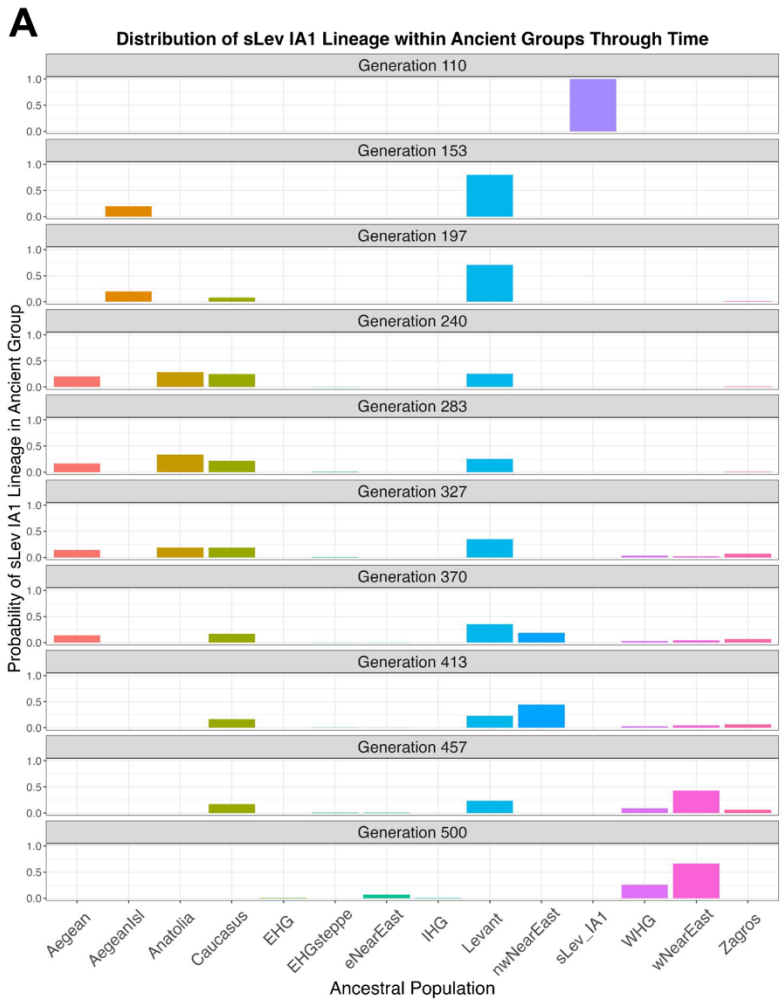
where c_x corresponds to the probability two lineages sampled in the Target population have a common ancestor before the time of admixture (Peter 2016).

By conditioning on simulations whose demographic parameters result in a negative f_3-statistic condition (i.e the value of EQ:1 left hand side (LHS) must be less than the value of the right-hand side (RHS)), we show that demographic Models with admixture to the source (S1) population from the iS2R2 branch (Models 3 and 4) result in the largest type II error rate (percent of simulations with f_3(Target; S1, S2) Z-score > -3 for Models 1, 2, 3, and, 4 are 34%, 30%, 44%, and 43%, respectively) (SI Figure S5A-D). As such, we show that the Models with a gene flow from the iS2R2 branch require, on average, a larger LHS to RHS difference (smaller ratio of LHS / RHS) for the negative f_3-statistic condition to generate significance (median EQ:1 LHS / RHS ratio for f_3(Target; S1, S2) Z-scores < -3 across Models 1, 2, 3 and, 4 are 0.158, 0.128, 0.085, and 0.078, respectively). Given a substantial period of independent drift between the ancestral split of the Sources and the time of admixture is a prominent factor in f_3-statistic negativity, this power reduction appears to be principally driven by the increase in the differences between T_{admix} / T_2 caused by admixture between the ancestral source lineages. Importantly, all the demographic effects on f_3-statistics power described above are magnified when selecting the wrong source pairs (SI Figure S5B-D).

qpAdm model ranking by *P*-value

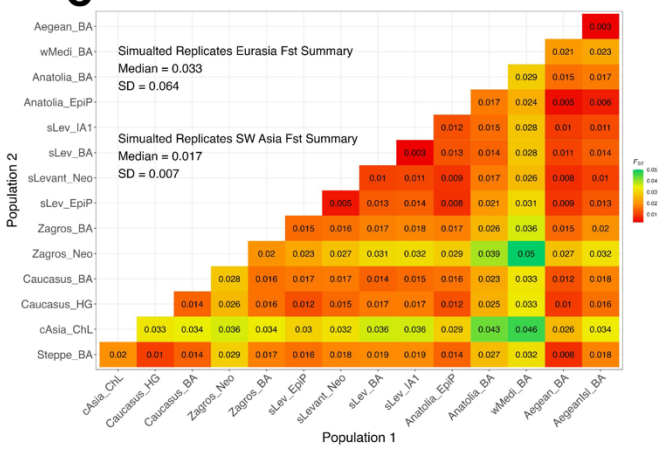
A common application of qpAdm, and by extension qpWave, is ranking model performance via *P*-values (van de Loosdrecht *et al.* 2018; Oliveira *et al.* 2022; Lazaridis *et al.* 2022a; Taylor *et al.* 2023; Moots *et al.* 2023). We evaluated the use of *P*-values for the relative ranking of qpAdm models by assessing how frequently each of the single and two-source qpAdm models had the largest, second-largest, third, and fourth-largest *P*-values for each of the 5k simulations. Across all demographic Models, the true qpAdm model significantly outperformed all other qpAdm models by having the largest *P*-value in more than 60% of the simulations (SI Table ST1A-D). Additionally, we found that both the relative ranking and frequency of *P*-values reflected the underlying demography and frequency of plausible models described above (Figure 4). Under Model 1, the S1+R2 and S2+R1 models had the largest *P*-value with about equal frequency (0.127 and 0.137, respectively), whereas, under demographic Models 2-4, the S1+R2 qpAdm model has the largest *P*-value approximately 10x more frequently and is the second best performing of all qpAdm models (SI Table ST1A-D).

Figure 6



B

Ancestral Population	Population	Sampling Date (Generations)	qpAdm Function
Levant	sLev_IA1	106	Target
Levant	sLev_IA2	101	Target
Levant	sLev_IA3	86	Target
Levant	sLev_Hist1	61	Target
Levant	sLev_Hist2	31	Target
Levant	sLev_Hist3	11	Target
Mbuti	Mbuti	0	Right Fixed
BHG	Balkan_HG	340	Right Fixed
Zagros	Zagros_Neo	360	Right Fixed
Levant	sLevant_Neo	360	Right Fixed
Caucasus	Caucasus_HG	395	Right Fixed
Levant	sLev_EpiP	430	Right Fixed
nwNearEast	Anatolia_EpiP	440	Right Fixed
nAfr	nAfr_EpiP	490	Right Fixed
nEurasia	nEur_HG	830	Right Fixed
OOA	UstIshim	1550	Right Fixed
wMediterranean	wMedi_BA	120	Source Rotate
AegeanIsl	AegeanIsl_BA	132	Source Rotate
Aegean	Aegean_BA	145	Source Rotate
Levant	sLev_BA	150	Source Rotate
Caucasus	Caucasus_BA	155	Source Rotate
Steppe	Steppe_BA	160	Source Rotate
Anatolia	Anatolia_BA	165	Source Rotate
Zagros	Zagros_BA	170	Source Rotate
cAsia	cAsia_ChL	175	Source Rotate
WHG	WEHG	265	Source Rotate
EHG	EEHG	295	Source Rotate



(A) Barplots showing probabilities of encountering a lineage found in the "*sLev IA1*" group in other simulated ancestral populations (only presenting populations with non-negative probabilities). The ancestral populations are those from which we sampled and correspond to the first column in B. (B) A table of the sampled populations used in qpAdm analysis and the and the ancestral populations they split from (corresponding to ancestral populations in A). An F_{ST} matrix (C) for the sampled simulated populations is also shown.

Going complex: Admixture Inference under Complex Human History in aDNA Research

We expanded our evaluation of admixture inference from simple topologies to a demographic model and data distribution that reflects the real-world complexities of both Eurasian human history and aDNA conditions. We framed this by simulating an archaeogenetic hypothesis on the origin of migrants to the southern Levant at the beginning of the Iron Age (the so-called Sea Peoples migration). While our demographic model and parameters are informed by the aDNA and population genetic literature, we stress that it is not designed to represent true human history, nor a proposal of the likely events associated with the Sea Peoples migration. Rather, its function is solely to capture some of the complexities surrounding the dynamics connecting populations in the historical period such as low divergence between candidate source populations, complexity of ancestral

population relationships, and sampling recently after admixture event. As such, it provides us a framework from which we can evaluate the behavior and limitations of admixture inference from aDNA.

In total, we model 59 populations and 41 pulse admixture events (Figure 6A) which are all described and referenced in Supplementary File SF1. A brief summary of the model scaffold follows. The oldest split in the demography is the separation of East and Central African ancestral populations at 5,172 generations before present (Hollfelder *et al.* 2021). We model an out-of-Africa (OOA) population as separating from the East African lineage at 3,303 generations (Kamm *et al.* 2020; Marchi *et al.* 2022), and from the former lineage split East Eurasian, North Eurasian, West European Hunter-Gatherer (WEHG), and ancestral Near Eastern lineages (Kamm *et al.* 2020; Marchi *et al.* 2022). Two meta Near Eastern lineages, Eastern and Western Near East, split from the ancestral Near Eastern lineage (Marchi *et al.* 2022). The Levant lineage, from which the target southern Levant IA population ("sLev_IA1") largely descends, splits from the "Western Near East" lineage at 483 generations (Lazaridis *et al.* 2016; Broushaki *et al.* 2016; Marchi *et al.* 2022). We model the formation of a "Northwestern Near East" lineage at 446 generations, from which the admixing source population "Aegean Island" ("AegeanIsl_BA") largely descends, as a mix of "Western Near East" (0.86) and WEHG (0.14) (Marchi *et al.* 2022). The Target lineage, sLev_IA1, was modeled as a mixture of its ancestral population (southern Levant Bronze Age, "sLev_BA") and the AegeanIsl_BA population (admixture fraction = 0.2) at generation 111 before present. We sampled the Target population five generations post-admixture (Supplementary File SF1). To assess the influence of post-admixture drift on admixture inference, we modeled successive step-wise splits from the Target lineage and sampled them 10, 25, 50, 80, and 100 generations post the original admixture event. From our simulated lineages, we sampled data representing the Mbuti present-day population, and 20 ancient Eurasian and African populations that reflect empirical ancient groups present in many Southwest Asian aDNA analyses (Figure 6B). For all populations, we sampled 10 individuals and in all downstream analyses we defined the pairing of sLev_BA+AegeanIsl_BA as the true model and all others as false models. As above, we consider plausible models to have a *P*-value ≥ 0.05 and admixture weights between zero and one ([0:1]).

Data generation

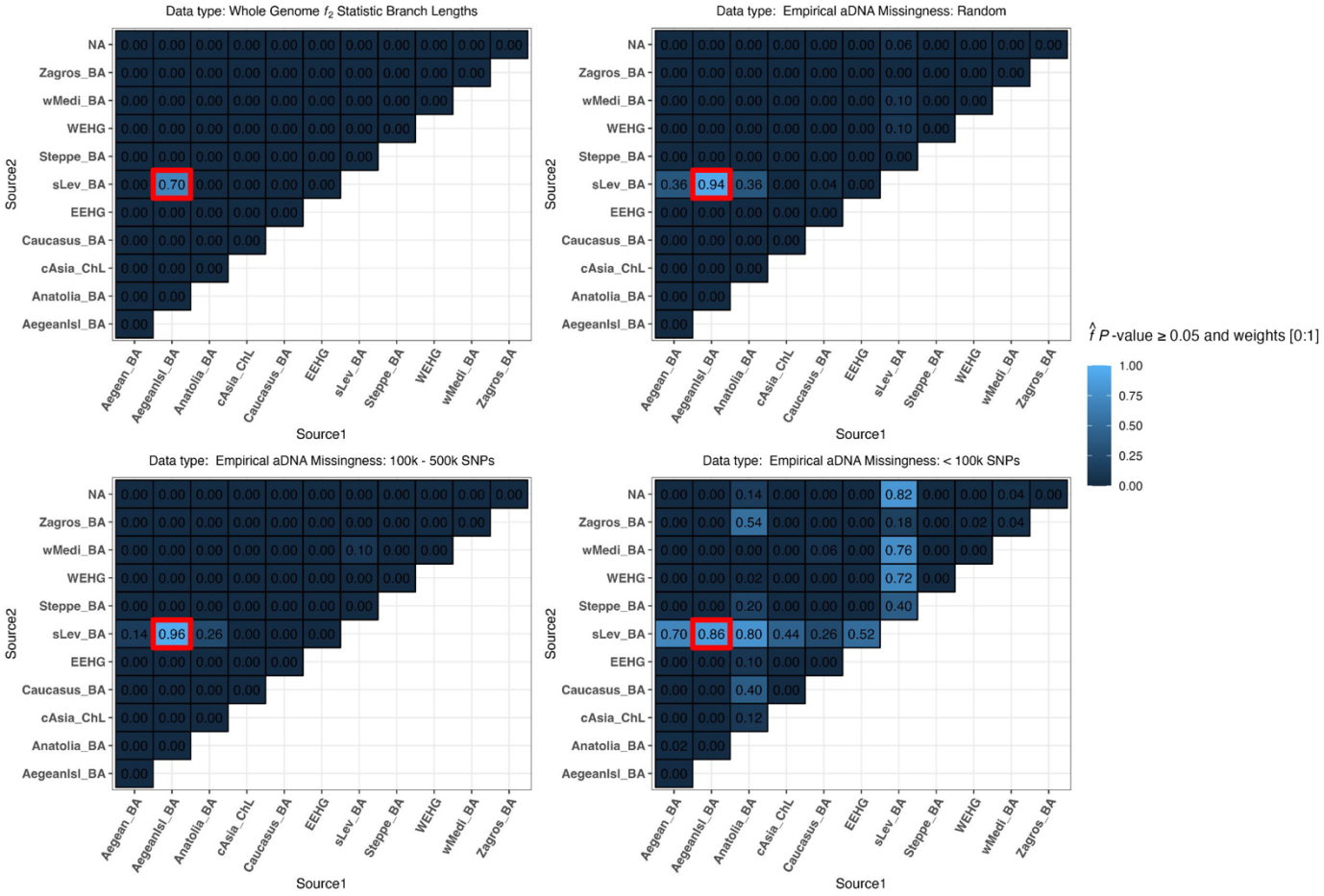
We configured the Eurasian demographic Model (Supplementary File SF1) using the Demes graph format (Gower *et al.* 2022) and converted it to an msprime demography object through the demography.from_demes() function. We simulated 50 whole-genome (*L*~ 2875 Mbp) replicates using sequence lengths and recombination rates of chromosomes 1-22 following the HomSap ID from the stdpopsim library (Adrion *et al.* 2020) and separated each chromosome with a log(2) recombination rate following msprime manual guidelines (Nelson et al. 2020, Baumdicker et al. 2022). The first 25 generations into the past were simulated under the Discrete Time Wright-Fisher (DTWF) model (Nelson *et al.* 2020), and under the Standard (Hudson) coalescent model until the sequence MRCA. We applied mutations to the simulated tree sequence at a rate of 1.29e-08 (Jónsson *et al.* 2017) using the Jukes-Cantor mutation model (Jukes and Cantor 1969). From the mutated tree sequence, we generated Eigenstrat files through the tskit v.0.5.2 TreeSequence.variants() function which were passed to

589 custom R scripts to generate realistic aDNA conditions such as filtering on bi-allelic sites, adding ascertainment
590 bias, downsampling to 1,233,013 SNPs (1240k capture), and for the simulated ancient individuals, generating
591 pseudo-haploid data with high missing rate (Github Repo:https://github.com/archgen/complex_demog_sims.git).
592
593 We configured the SNP ascertainment bias scheme replicating the general principles of the Human Origins
594 array (Patterson *et al.* 2012). See also (Flegontov *et al.* 2023) for an overview of effects of this type of
595 ascertainment on *f*-statistics and related methods. In the Eurasian demography, we defined separate lineages
596 representing central European (CEU), East Asian (CHB), African (AFR), and South Asian (sAs) populations,
597 sampled a single individual from these lineages at the present, and retained biallelic sites that are heterozygous
598 in at least one of these individuals. We then downsampled the simulated data by randomly sampling 1,233,013
599 SNP loci. For the simulated ancient samples (Figure 6B), we randomly assigned one of the two alleles as
600 homozygous at simulated heterozygous positions mimicking what is commonly performed for low- and medium-
601 coverage aDNA (Schuenemann *et al.* 2017). In addition, we added missing data by assigning to each ancient
602 individual an empirical missingness distribution from a randomly selected ancient individual within the AADR
603 v.52.2 (Mallick *et al.* 2023), which we filtered by removing related and contaminated ancient individuals and
604 restricted to individuals from Southwest Asia (see Supplementary File SF2 for the list of aDNA individuals).
605 Amongst the Target populations, this resulted in a range of missingness within each replication (the median
606 standard deviations of the population missingness across the replicates ranged from 0.04 to 0.14), with the
607 average proportion of missingness across the 50 replicates ranging from a minimum of 49% for the sLev_IA3
608 population to 89% in the sLev_IA1 population (resulting in a range of approximately 130,656 to 627,948 useful
609 SNPs, respectively) (SI Table ST2). We observe similar degrees of missingness for the other ancient
610 populations included in the qpAdm rotation analysis (SI Table ST2). We generated two additional missing data
611 subsets following the method above, whereby the AADR individuals were filtered to contain only low (SNPs <
612 100k), or medium coverage (100k < SNPs < 500k) from samples across Eurasia, resulting in 50 whole-genome
613 replicate simulations with three different degrees of missingness.
614
615 From the simulated aDNA we computed rotating qpAdm analyses with the ADMIXTOOLS2 software (Maier *et*
616 *al.* 2023) using parameters typical of empirical aDNA workflows such as “allsnps = TRUE” (using all SNPs
617 available for calculating each individual f_4-statistic), and 5 Mbp windows for calculating standard errors of f_4-
618 statistics with the jackknife procedure. We configured the qpAdm analysis protocol in the following way: the
619 most ancient groups are fixed in the right-group position and younger candidate populations are rotated
620 between the left and right-group positions (Narasimhan *et al.* 2019; Lazaridis *et al.* 2022a). The Mbuti
621 population was fixed in the first position in all qpAdm analyses, along with nine deeply divergent Eurasian and
622 African populations fixed in the right group. We then rotated nine simulated Bronze Age and Chalcolithic, and
623 two European hunter-gatherer populations (Figure 6B) between the left and right-group positions resulting in a
624 total of 11 single-source models, and 66 two-source models.
625

To evaluate the impact of aDNA conditions on admixture inference under the Eurasian human demography, we generated f_2 - and F_{ST} - statistic matrices directly from the tree sequence without mutations through tskit v.0.5.2 with parameters “Mode=branch”, and “span_normalise=True”, using 5 Mbp windows. The resulting f_2 - statistics matrix was used to compute rotating qpAdm analyses using the same sample set as input for the aDNA application, and admixture f_3-statistic in the ADMIXTOOLS2 software (Maier *et al.* 2023).

Our simulations resulted in expected levels of population divergence given empirical observations with a median pairwise F_{ST} of 0.03 between all Eurasian populations and 0.017 amongst the Southwest Asian populations (Figure 6C). A pairwise F_{ST} matrix computed on the first replicate shows expected genetic affinities amongst the analysis populations (Figure 6C). We used the tskit lineage_probabilities() function to further assess the relationship between the Target and analysis populations by tracking the location of lineages sampled from the Target through time amongst the remaining simulated demographic lineages (Figure 6A). The results show that between the youngest and oldest populations included in the qpAdm analysis, lineages from the Target population are principally found in the Levant, Aegean, AegeanIsl, Anatolia, and Caucasus ancestral groups (Figure 6A).

Figure 7



Heatmaps of the proportion of replicates with plausible *P*-value ≥ 0.05 and weights [0:1] for the complex demography (Aegean Island admixture to southern Levant) for the target population "sLev IA1". The red box represents the most optimal true model. Results are presented for four datasets: f_2-statistics calculated on whole-genome branch lengths and the three datasets with varying SNP missing rates.

qpAdm performance and aDNA data quality

Consistent with the results previously shown by (Harney *et al.* 2021), the degree of data missingness appears to be one of the primary factors influencing the performance of rotating qpAdm. Below, we adopt the term "coverage" to represent the proportion of SNPs non-missing. Thus, the aDNA missingness sampling condition of SNPs < 100k represents the lowest-coverage dataset, the random missingness sampling condition represents the medium-coverage dataset, and the missingness condition of 100k < SNPs < 500k represents the highest-coverage dataset. The lowest-coverage dataset produced the largest frequency of plausible single-source and two-source qpAdm models (Figure 7), resulting in the lowest average QTP, largest FPR and FDR, and an average QTP-binary of zero (SI Table ST3). The two higher-coverage aDNA sampled datasets resulted in very similar qpAdm performance with the highest-coverage dataset performing slightly better than the middle coverage dataset as it both rejects all single-source qpAdm models and has less total plausible qpAdm models (Figure 7), resulting in on average higher QTP and QTP-binary and the lowest FPR and FDR (SI Table ST3). Since the degree of missingness in the AADR random sampling scheme sometimes results in populations with more missingness than the lowest-coverage dataset (SI Table ST2), the relative performance of these missing data schemes highlights the importance of maximizing data coverage in all populations, not just the Target, for rejecting false qpAdm models.

Interestingly, we note that amongst the two higher-coverage datasets the plausible false qpAdm models are not arbitrarily selected as they descend from ancestral populations that are shown above to harbor the Target population lineages (Figure 6A). The single exception to this pattern is the simulated wMedi_BA population which, when only paired with the sLev_BA population, is plausible at 10% in each of the higher coverage datasets, and 76% in the lowest-coverage dataset (Figure 7). The simulated wMediterranean lineage, from which wMedi_BA descends, is modeled as receiving 6% admixture from the true source population 29 generations before the formation of the Target population (Supplementary File SF1). This demonstrates that correlations in allele frequencies between populations driven by admixture from a shared ancestral source, in addition to shared genetic drift, can result in false positive qpAdm results.

Interestingly, the qpAdm rotation analysis on the simulated whole-genome branch-length f_2-statistic rejected all false qpAdm models and classified the true model plausible in 70% of the replicates and rejected it in 30% of the replicates (Figure 7). We ran a receiver operating characteristic curve (ROC) analysis where we varied the *P*-value between zero and one to assess the relationship between *P*-value thresholds and qpAdm performance

as measured by the true positive (TP) and false positive (FP) rates. In calculating the ROC, we constrained the qpAdm models to have plausible admixture weights [0:1] and performed each calculation on 3,000 *P*-value thresholds between zero and one. These results revealed for the whole-genome branch-length f_2-statistic dataset, the qpAdm TPR converges to 100% with *P*-values greater than 1×10^{-3} and the FPR does not increase until the *P*-value reaches zero (SI Figure S6). All datasets with aDNA missingness exhibit a trade-off of co-varying increases/decreases in the FPR/TPR with changes to the *P*-value threshold (SI Figure S6) which we also observe in the distribution of *P*-values for qpAdm models with plausible admixture weights (SI Figure S7). Importantly, both higher-coverage aDNA datasets have greater than 89% TPR and less than 1% FPR with a *P*-value threshold of 0.1, suggesting an additional strategy for increasing qpAdm accuracy (SI Figure S6).

qpAdm model ranking by *P*-value in ancient DNA under complex demography

We evaluated the accuracy of determining the best-fitting qpAdm model by ranking them by their *P*-values given the admixture complexity of the simulated Eurasian demography. Under the higher coverage datasets, the true model has the largest *P*-value in more than 90% of the simulation replicates and has the largest *P*-value in 100% of the replicates using the highest-coverage dataset (SI Table ST4). However, caution should be applied to ranking qpAdm by *P*-values in datasets with low coverage as in our lowest aDNA coverage dataset, the true model has the largest *P*-value in only 16% of the replicates, second to the false model of sLev_BA+Anatolia_BA (SI Table ST4). The observation of the sLev_BA+Anatolia_BA and sLev_BA+Aegean_BA source combinations as the alternative qpAdm models that possessed the largest qpAdm *P*-value in at least one replicate demonstrates the difficulty in rejecting closely related candidate sources (F_{ST} between the AegeanIsl_BA or Aegean_BA and Anatolia_BA populations ~ 0.003 and 0.017, respectively). Nonetheless, the sLev_BA population is consistently paired with alternative sources in the most frequent qpAdm models with the largest *P*-values, suggesting that the identification of overrepresented populations in high-ranking qpAdm models is a suitable heuristic to determine a likely true source regardless of the degree of data missingness (SI Table ST4).

Generations since admixture and qpAdm performance and accuracy

We also explored the effect of post-admixture drift on qpAdm performance. Importantly, across all descendent Target populations and degrees of data missingness, we observe no significant trend in qpAdm performance or admixture weight accuracy (SI Figure S8). Under the whole-genome branch-length f_2-statistic dataset, the admixture weight S.E. appears to increase with increasing generations since admixture, however, it has no significant impact on accuracy or precision of admixture weight estimates (SI Figure S8). As expected, we observe the largest estimated admixture weight S.E. and delta-alpha values in the lowest-coverage dataset, with between 0.12 and 0.20 SD on delta-alpha (SI Figure S8). As such, caution should be given to interpreting admixture proportions from datasets of low coverage.

qpAdm plausibility criteria

We evaluated the impact of the different qpAdm plausibility criteria described in the simple demography section on our complex demographic aDNA simulations. In contrast to the simple demographic simulations, the introduction of the 2 S.E. constraint on admixture weight estimates consistently either reduced or maintained the FPR for all aDNA missingness conditions and either reduced or maintained the FDR in all but the lowest-coverage datasets (SI Table ST3). Of note is that each aDNA missingness dataset has a different plausibility criterion that maximizes its QTP, making the selection of single plausibility criteria to maximize QTP infeasible. We do, however, observe for all datasets, the lowest error rates (FDR and FDR) with the co-criteria of rejection of all single-source models, *P*-value ≥ 0.05, and 2 S.E. constraint on the admixture weight estimates, albeit with greater than 0.98 FDR in the lowest coverage dataset (SI Table ST3). The plausibility criteria of *P*-value ≥ 0.05 and admixture weights [0:1] results in the smallest FDR in the lowest-coverage dataset. In empirical studies with low coverage aDNA samples, this may represent the most optimal plausibility criteria as it minimizes the frequency of type II errors as evidenced by the largest QTP value (SI Table ST3). The distribution of *P*-values for models with plausible admixture weights [0:1] (SI Figure S7), and the ROC curve analysis (SI Figure S6) shows that increasing the *P*-value threshold for the low-coverage dataset does not result in a significant reduction in the FP rate without penalizing the TP rate (SI Figure S6).

Importantly, we observe no impact from the use of significant admixture f_3-statistics as an additional plausibility criterion to increase qpAdm model inference accuracy as all pairwise combinations of qpAdm sources were not significant regardless of data quality (SI Figure S9). As described above, the power for the detection of admixture from f_3-statistics is strongly influenced by the underlying population demography and divergence of the candidate sources from the true admixing populations. Our simple demography simulations showed that both gene flow between source lineages, and increased divergence of the candidate source population from the true admixing source reduced the f_3-statistics power. Under the Eurasian demography, the two source groups, Levant and Aegean, undergo recent bi-directional gene flow after the split from their most recent common ancestral population.

We computed the f_3-statistics negativity condition (EQ:1) for both the split-time of the Levant and Aegean sources and the date of admixture for all Target populations. As expected, we observe an increase in f_3-statistic estimates with increasing generations since admixture (SI Figure S9). Moreover, the estimated f_3-statistic negativity appears to conform closer to the f_3-statistics negativity condition (EQ:1) when computed using the date of most recent bi-directional admixture between the Levant and Aegean sources than the their split time (SI Figure S9). This further supports the impact of admixture between source lineages on decreasing the power of the admixture f_3-statistic and highlights the importance of utilizing f_3-statistic estimates as confirming plausible

qpAdm models rather than rejecting false models, similar to how they were originally proposed (Patterson *et al.* 2012).

Data availability

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, tables, and supplementary materials. Both simple demographic Model and complex Eurasian model simulations were written in Snakemake pipelines to facilitate reproducibility and can be accessed via our GitHub repository https://github.com/archgen/complex_demog_sims. Supplemental figures available in Supplementary Material PDF:

Discussion

The qpAdm software has become one of the hallmark methods in archaeogenetic analyses for reconstructing admixture histories of ancient populations (see Lazaridis *et al.* 2016, 2022a; b; Skoglund *et al.* 2017; Mathieson *et al.* 2018; Harney *et al.* 2018; Narasimhan *et al.* 2019; Antonio *et al.* 2019; Marcus *et al.* 2020; Fernandes *et al.* 2020; Wang *et al.* 2020, 2021; Ning *et al.* 2020; Yang *et al.* 2020; Carlhoff *et al.* 2021; Papac *et al.* 2021; Librado *et al.* 2021; Sirak *et al.* 2021; Patterson *et al.* 2022; Changmai *et al.* 2022a; b; Bergström *et al.* 2022; Maróti *et al.* 2022; Lee *et al.* 2023). This is due in part to its modest computational requirements, use of allele frequency data, and minimal model assumptions (Haak *et al.* 2015; Harney *et al.* 2021). The primary motivation for this work is addressing its applicability, performance, and limits in reconstructing admixture histories under challenging scenarios that emerge when reconstructing population dynamics within the historical period. Such conditions range from identifying the true source population amongst minimally differentiated candidates, and potential biases that may arise from sources that are admixed and ancestrally connected through complex demographies. It also may involve dealing with short intervals between the admixture event of interest and the ancient sample. Additionally, we sought to determine how these challenges are impacted by missing data typical of aDNA conditions. We addressed these questions through simulations of both simple admixture-graph-like demographies exploring a wide parameter space, and whole-genome simulations of an admixture-graph-like demography that reflects the inferred complexity of Eurasian population history.

It is important to acknowledge that our study configures human demography as a series of discrete population splits and pulse admixture events, each separated by periods of independent genetic drift (that is why these simulations are termed “admixture-graph-like”). Thus, if the distribution of human settlements across the ancient landscape aligns more closely with temporally evolving stepping-stone models, the interpretations drawn from our study may lose some of their significance. Also of note is that all of our demographic models adhere to the

fundamental assumptions of qpAdm (Harney *et al.* 2021): 1) there are no gene flows connecting lineages private to candidate source populations (after their divergence from the true admixing populations) and "right-group" populations, and 2) there are no gene flows from the fully formed Target lineage to "right-group" populations (Harney *et al.* 2021). It is crucial to recognize that these assumptions might be frequently violated when investigating demographic history in the historical period and beyond it, leading to false rejections of true simple models. In turn, these prompt researchers to test more complex models which often satisfy qpAdm model plausibility criteria but are misleading when subjected to historical interpretation (Yüncü *et al.* 2023). If stringent sampling criteria, as outlined in our companion paper (Yüncü *et al.* 2023), are not diligently followed, these violations are shown to pose substantial challenges to the effective use of qpAdm in demographic inference.

Our simple demographic simulation results show that qpAdm converges on its maximum QTP as the median pairwise F_{ST} of the sample set approaches ~ 0.005 - 0.008 (Figure 3E-H), well within the diversity expected of historical period populations (Figure 1B). However, we find a much larger level of population divergence is required, with a median pairwise F_{ST} exceeding 0.015, to simultaneously reject all false models and identify the true model with a probability greater than 30% (Figure 3I-L). This finding suggests that highly specific archaeogenetic hypotheses that require the sole identification of the correct model may currently lie beyond the capabilities of qpAdm given the prevailing data conditions. Importantly however, within the set of models considered plausible by qpAdm under both the simple admixture simulations and Eurasian complex simulations, we consistently observe that one of the true sources is included in those most frequently accepted models, irrespective of the degree of population divergence or levels of data missingness (Figure 4A-D & Figure 7).

When it comes to distinguishing between closely related cladal populations, such as the differentiation between S1 and R1 or S2 and R2 in our simple admixture simulations, our results suggest that qpAdm exhibits heightened discriminatory power when these closely related cladal populations have diverged on the order of F_{ST} > ~ 0.002 - 0.004 (SI Figure S4). A similar result emerges from our complex Eurasian demographic simulations. For instance, the candidate source Aegean_BA, which is modeled as having recently split from the true source AegeanIsl_BA, is differentiated at a median F_{ST} of 0.003 and is frequently included in plausible qpAdm models at all levels of data missingness (Figure 7). However, it's worth noting that in the complex Eurasian demographic simulations, population divergence alone does not exclusively determine the probability of a false source appearing in a plausible qpAdm model. For instance, we frequently observe the Anatolian_BA population in plausible qpAdm models (Figure 7) whilst it is both approximately equally divergent from the true sources (median F_{ST} Aegean_BA = 0.016 and sLev_BA = 0.015) (Figure 6C). This is likely driven by demographic factors analogous to the conditions of simple demographic Models C-D (Figure 3 C-D) whereby the Eurasian demographic model includes bi-directional gene flow between the ancestral Levant and Anatolian populations (30% Levant to Anatolia, and 40% Anatolia to the Levant in generations 305 and 224, respectively)

resulting in a substantial likelihood that lineages from the Target population are present within the Anatolian population (Figure 6A).

A key discovery with relevance for archaeogenetic research in regions with complex migration histories is that introducing admixture into the source population (but not violating the topological assumptions described above) can notably improve qpAdm's performance, especially when considering conditions that resemble the typical levels of divergence observed during the historical period. Notably, the phylogenetic origin of the ancestral admixture differently impacts qpAdm accuracy (FPR and FDR) and performance (QTP). For instance, when the gene flow originates from an outgroup to the Target, true Sources, and candidate source populations, it yields the highest average QTP performance and lowest FDR among all demographic models. In contrast, we observe lower FP rates under demographic Models that include admixture between sources (Model 3) than from an outgroup (Model 2), and the lowest FP rate when both ancestral source admixture events occur (Model 4) (Table 1). Overall, this trend appears to be primarily driven by the increased differential relatedness between left and right-set qpAdm populations, irrespective of the decrease in average population divergence.

This observation is consistent with theoretical expectations regarding the way qpAdm uses *P*-values to reject candidate models (Haak *et al.* 2015; Harney *et al.* 2021). When the Target and right-group populations share genetic drift distinct from the shared ancestry between the Target and the putative left-group sources, this will result in the rejection of the left-group sources as an admixture model of the Target population given a certain *P*-value threshold. As such, the ancestry inherited by the Target from a source that is itself admixed increases the number of populations that it uniquely shares drift with. This is evident in the increased power to reject the false R1+S2 qpAdm models with the introduction of admixture to the S1 ancestral source lineage (Figure 4A-D). Consequently, these observations underscore the importance in empirical aDNA studies of pre-screening qpAdm right-groups to optimize genetic differentiation and differential relatedness with potential source populations for maximizing qpAdm performance, as originally proposed in (Haak *et al.* 2015).

We also note that as long as the correct source populations are chosen, usage of source populations with complex admixture history does not introduce bias in the estimation of admixture weights (Figure 5). However, we do observe a reduction in accuracy when admixture to a source lineage occurs from another source branch, in contrast to an outgroup (Figure 5). Most notably, the selection of source populations appears to be more critical for accurately estimating admixture contributions. We observed a significant bias towards the population closest to the true source, leading to an overestimation of admixture proportions for this population (Figure 5). This phenomenon is present in all simple demographic models but appears to be more pronounced in models with ancestral admixture to the source (Figure 5). In cases where both populations are equidistant from the true admixing sources, the bias is only evident in models that include an admixture event between source branches (Figure 5).

We have observed that two additional criteria significantly enhance the accuracy of qpAdm model inference. The first involves considering two-source (admixture) qpAdm models only when all single-source models are rejected. The second criterion involves deeming these models as plausible when the source pairs generate a significantly negative admixture f_3-statistic (Z-score < -3). While these criteria have proven effective in reducing bias (FPR and FDR) across a wide range of demographic parameters, in empirical studies it is crucial to assess the anticipated parameters of each demographic model being evaluated before applying these criteria universally. This is because certain demographic conditions can increase the FDR. For example, we observe an increase of plausible single-source models when the admixture weight is close to 1 (SI Figure S10), which, if requiring all single-source models to be rejected, will result in the more frequent false rejection of the true model.

As for the criterion of a significant admixture f_3-statistic, under conditions where there is only a short period between the split of the admixing source populations and their admixture to form the Target, and when the admixture proportions deviate significantly from 0.5, the power of the f_3-statistic to detect an admixture event diminishes, increasing type II errors (SI Figure S5). We observe this scenario in our Eurasian demographic simulations where admixture between the ancestral sources after their split decreased the power of the admixture f_3-statistic, resulting in a 100% type II error rate (SI Figure S9). Moreover, the simple simulations reveal this effect is exacerbated by the divergence between the tested candidate population and the true admixing source, making the conditions for negativity of the f_3-statistic even more stringent (SI Figure S5). Therefore, we suggest that the f_3-statistic should be used as a confirmation and ranking tool for plausible qpAdm models, rather than as a criterion for rejecting them (i.e. favoritism is given to plausible models with a significant f_3-statistic over those without). This aligns with the original interpretative guidance when using the f_3-statistic as a formal admixture test (Patterson *et al.* 2012).

With respect to the procedure of ranking feasible qpAdm models based on their *P*-values, the initial suggestions from Harney et al. in 2021 raised a notable concern and advised against *P*-value ranking to identify the best model. Their argument is based on the observation that *P*-values under false models closely related to the true model are almost uniformly distributed, and that in cases when multiple models are plausible any one false model could easily have a larger *P*-value than the true model (Harney *et al.* 2021). However, our findings from simple admixture demographic simulations show that in over 60% of the simulations, the true model has the largest *P*-value (SI Table ST1). Similarly, in the complex Eurasian simulations under the condition of high coverage, we found that in over 90% of the replications, the true model had the largest *P*-value (SI Table ST4). In both simple admixture and complex Eurasian simulations, we also found that the relative ranking of *P*-values accurately reflected how closely a false model represented the true ancestry of the Target population. Therefore, provided an empirical dataset has good coverage, we propose that ranking qpAdm models by *P*-values can offer valuable additional information for determining the true model.

A surprising finding from the complex Eurasian demographic simulations was when performing the qpAdm rotation analysis with f_2-statistics computed from the whole-genome branch lengths we obtained a 100% true positive rate and a 0% false positive rate with a *P*-value threshold of 0.001 (SI Figure S6). This outcome underscores the remarkable potential inherent in the principles underlying qpAdm, while also highlighting the constraints imposed by data scale. In light of this, we suggest there is room for significant enhancements in the qpAdm protocol from methods that can leverage more accurate estimations of f_2-statistics. Possible avenues for this improvement could be developing innovative techniques for extracting information from ancestral recombination graphs (ARG) within the context of aDNA, or conditioning on the site frequency spectrum (SFS) for the enrichment of rare alleles. As such, it is clear from these results that further improvements would make qpAdm a powerful tool for accurately reconstructing the genetic histories of populations under the most complex scenarios.

Acknowledgments

We are grateful for the advice and helpful suggestions from Yassine Souilmi, Raymond Tober, and Bastien Llamas at an initial stage of the project. All simulations were performed on the computational infrastructure at the Pennsylvania State University's Institute for Computational and Data Sciences' Roar supercomputer. We note that our content is solely the responsibility of the authors and does not necessarily represent the views of the Institute for Computational and Data Sciences.

Funding

C.D.H. and M.P.W were funded by the National Institute of Health under award number R35GM146886. P.F. was supported by the Czech Science Foundation (project no. 21-27624S led by P.F.), the Czech Ministry of Education, Youth and Sports (program ERC CZ, project no. LL2103), the John Templeton Foundation (grant no. 61220 to David Reich), and by a gift from Jean-Francois Clin.

Conflicts of interest

None declared.

Literature cited

Adrion J. R., C. B. Cole, N. Dukler, J. G. Galloway, A. L. Gladstein, *et al.*, 2020 A community-maintained standard library of population genetic models.

Agranat-Tamir L., S. Waldman, M. A. S. Martin, D. Gokhman, N. Mishol, *et al.*, 2020 The Genomic History of the Bronze Age Southern Levant. Cell 181: 1146–1157.e11.

Alexander D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19: 1655–1664.

Antonio M. L., Z. Gao, H. M. Moots, M. Lucci, F. Candilio, *et al.*, 2019 Ancient Rome: A genetic crossroads of Europe and the Mediterranean. Science 366: 708–714.

Arning N., and D. J. Wilson, 2020 The past, present and future of ancient bacterial DNA. Microb Genom 6.

Ávila-Arcos M. C., M. Raghavan, and C. Schlebusch, 2023 Going local with ancient DNA: A review of human histories from regional perspectives. Science 382: 53–58.

Barros Damgaard P. de, R. Martiniano, J. Kamm, J. V. Moreno-Mayar, G. Kroonen, *et al.*, 2018 The first horse herders and the impact of early Bronze Age steppe expansions into Asia. Science 360.

Bartash V., 2020 The Early Dynastic Near East, in Oxford University Press.

Baumdicker F., G. Bisschop, D. Goldstein, G. Gower, A. P. Ragsdale, *et al.*, 2021 Efficient ancestry and mutation simulation with msprime 1.0. Genetics 220: iyab229.

Bergström A., D. W. G. Stanton, U. H. Taron, L. Frantz, M.-H. S. Sinding, *et al.*, 2022 Grey wolf genomic history reveals a dual ancestry of dogs. Nature 607: 313–320.

Boyle K., and C. Renfrew, 2000 Archaeogenetics : DNA and the population prehistory of Europe, in *Published in **2000** in Cambridge by McDonald institute for archaeological research*, Cambridge : McDonald Institute for

943 Archaeological Research.

944 Broushaki F., M. G. Thomas, V. Link, S. López, L. van Dorp, *et al.*, 2016 Early Neolithic genomes from the
945 eastern Fertile Crescent. Science 353: 499–503.

946 Brunson K., and D. Reich, 2019 The Promise of Paleogenomics Beyond Our Own Species. Trends in Genetics
947 35: 319–329.

948 Carlhoff S., A. Duli, K. Nägele, M. Nur, L. Skov, *et al.*, 2021 Genome of a middle Holocene hunter-gatherer from
949 Wallacea. Nature 596: 543–547.

950 Changmai P., K. Jaisamut, J. Kampuansai, W. Kutanan, N. E. Altınışık, *et al.*, 2022a Indian genetic heritage in
951 Southeast Asian populations. PLoS Genet. 18: e1010036.

952 Changmai P., R. Pinhasi, M. Pietrusewsky, M. T. Stark, R. M. Ikehara-Quebral, *et al.*, 2022b Ancient DNA from
953 Protohistoric Period Cambodia indicates that South Asians admixed with local populations as early as 1st-
954 3rd centuries CE. Sci. Rep. 12: 22507.

955 Clemente F., M. Unterländer, O. Dolgova, C. E. G. Amorim, F. Coroado-Santos, *et al.*, 2021 The genomic
956 history of the Aegean palatial civilizations. Cell 184: 2565–2586.e21.

957 De Schepper S., J. L. Ray, K. S. Skaar, H. Sadatzki, U. Z. Ijaz, *et al.*, 2019 The potential of sedimentary ancient
958 DNA for reconstructing past sea ice evolution. ISME J. 13: 2566–2577.

959 Durand E. Y., N. Patterson, D. Reich, and M. Slatkin, 2011 Testing for ancient admixture between closely
960 related populations. Mol. Biol. Evol. 28: 2239–2252.

961 Elise Lauterbur M., M. I. A. Cavassim, A. L. Gladstein, G. Gower, N. S. Pope, *et al.*, 2022 Expanding the
962 stdpopsim species catalog, and lessons learned for realistic genome simulations. bioRxiv
963 2022.10.29.514266.

964 Fernandes D. M., A. Mittnik, I. Olalde, I. Lazaridis, O. Cheronet, *et al.*, 2020 The spread of steppe and Iranian-
965 related ancestry in the islands of the western Mediterranean. Nat Ecol Evol 4: 334–345.

966 Flegontov P., U. Işıldak, R. Maier, E. Yüncü, P. Changmai, *et al.*, 2023 Modeling of African population history
967 using f-statistics is biased when applying all previously proposed SNP ascertainment schemes. PLoS
968 Genet. 19: e1010931.

969 Fu Q., C. Posth, M. Hajdinjak, M. Petr, S. Mallick, *et al.*, 2016 The genetic history of Ice Age Europe. Nature
970 534: 200–205.

971 Gower G., A. P. Ragsdale, G. Bisschop, R. N. Gutenkunst, M. Hartfield, *et al.*, 2022 Demes: a standard format
972 for demographic models. Genetics 222.

973 Green R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, *et al.*, 2010 A draft sequence of the Neandertal
974 genome. Science 328: 710–722.

975 Haak W., I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, *et al.*, 2015 Massive migration from the steppe was
976 a source for Indo-European languages in Europe. Nature 522: 207–211.

977 Haber M., M. Mezzavilla, Y. Xue, and C. Tyler-Smith, 2016 Ancient DNA and the rewriting of human history: be
978 sparing with Occam's razor. Genome Biol. 17: 1.

979 Haber M., C. Doumet-Serhal, C. Scheib, Y. Xue, P. Danecek, *et al.*, 2017 Continuity and Admixture in the Last
980 Five Millennia of Levantine History from Ancient Canaanite and Present-Day Lebanese Genome
981 Sequences. Am. J. Hum. Genet. 101: 274–282.

982 Haber M., J. Nassar, M. A. Almarri, T. Saupe, L. Saag, *et al.*, 2020 A Genetic History of the Near East from an
983 aDNA Time Course Sampling Eight Points in the Past 4,000 Years. Am. J. Hum. Genet.

984 Harney É., H. May, D. Shalem, N. Rohland, S. Mallick, *et al.*, 2018 Ancient DNA from Chalcolithic Israel reveals
985 the role of population mixture in cultural transformation. Nat. Commun. 9: 3336.

986 Harney É., N. Patterson, D. Reich, and J. Wakeley, 2021 Assessing the performance of qpAdm: a statistical tool
987 for studying population admixture. Genetics.

988 Harris A. M., and M. DeGiorgio, 2017 Admixture and Ancestry Inference from Ancient and Modern Samples

through Measures of Population Genetic Drift. Hum. Biol. 89: 21–46.

Hollfelder N., G. Breton, P. Sjödin, and M. Jakobsson, 2021 The deep population history in Africa. Hum. Mol. Genet. 30: R2–R10.

Jónsson H., P. Sulem, B. Kehr, S. Kristmundsdottir, F. Zink, *et al.*, 2017 Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature 549: 519–522.

Jukes T. H., and C. R. Cantor, 1969 CHAPTER 24 - Evolution of Protein Molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by Munro H. N. Academic Press.

Kamm J., J. Terhorst, R. Durbin, and Y. S. Song, 2020 Efficiently inferring the demographic history of many populations with allele count data. J. Am. Stat. Assoc. 115: 1472–1487.

Kelleher J., A. M. Etheridge, and G. McVean, 2016 Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLoS Comput. Biol. 12: e1004842.

Koptekin D., E. Yüncü, R. Rodríguez-Varela, N. E. Altınışık, N. Psonis, *et al.*, 2023 Spatial and temporal heterogeneity in human mobility patterns in Holocene Southwest Asia and the East Mediterranean. Curr. Biol. 33: 41–57.e15.

Kristiansen K., 2016 Interpreting Bronze Age Trade and Migration, pp. 154–180 in *Human Mobility and Technological Transfer in the Prehistoric Mediterranean*, Cambridge University Press.

Lazaridis I., D. Nadel, G. Rollefson, D. C. Merrett, N. Rohland, *et al.*, 2016 Genomic insights into the origin of farming in the ancient Near East. Nature 536: 419–424.

Lazaridis I., A. Mittnik, N. Patterson, S. Mallick, N. Rohland, *et al.*, 2017 Genetic origins of the Minoans and Mycenaeans. Nature 548: 214–218.

Lazaridis I., S. Alpaslan-Roodenberg, A. Acar, A. Açıkkol, A. Agelarakis, *et al.*, 2022a The genetic history of the Southern Arc: A bridge between West Asia and Europe. Science 377: eabm4247.

Lazaridis I., S. Alpaslan-Roodenberg, A. Acar, A. Açıkkol, A. Agelarakis, *et al.*, 2022b A genetic probe into the

ancient and medieval history of Southern Europe and West Asia. Science 377: 940–951.

Lee J., B. K. Miller, J. Bayarsaikhan, E. Johannesson, A. Ventresca Miller, *et al.*, 2023 Genetic population structure of the Xiongnu Empire at imperial and local scales. Sci Adv 9: eadf3904.

Librado P., N. Khan, A. Fages, M. A. Kusliy, T. Suchan, *et al.*, 2021 The origins and spread of domestic horses from the Western Eurasian steppes. Nature 598: 634–640.

Lipson M., P.-R. Loh, A. Levin, D. Reich, N. Patterson, *et al.*, 2013 Efficient moment-based inference of admixture parameters and sources of gene flow. Mol. Biol. Evol. 30: 1788–1802.

Lipson M., P.-R. Loh, N. Patterson, P. Moorjani, Y.-C. Ko, *et al.*, 2014 Reconstructing Austronesian population history in Island Southeast Asia. Nat. Commun. 5: 4689.

Liu Y., X. Mao, J. Krause, and Q. Fu, 2021 Insights into human history from the first decade of ancient human genomics. Science 373: 1479–1484.

Llamas B., E. Willerslev, and L. Orlando, 2017 Human evolution: a tale from ancient genomes. Philos. Trans. R. Soc. Lond. B Biol. Sci. 372.

Loosdrecht M. van de, A. Bouzouggar, L. Humphrey, C. Posth, N. Barton, *et al.*, 2018 Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations. Science 360: 548–552.

Maier R., P. Flegontov, O. Flegontova, U. Isildak, P. Changmai, *et al.*, 2023 On the limits of fitting complex models of population history to f-statistics.

Mallick S., A. Micco, M. Mah, H. Ringbauer, I. Lazaridis, *et al.*, 2023 The Allen Ancient DNA Resource (AADR): A curated compendium of ancient human genomes. bioRxiv 2023.04.06.535797.

Marchi N., L. Winkelbach, I. Schulz, M. Brami, Z. Hofmanová, *et al.*, 2022 The genomic origins of the world's first farmers. Cell 185: 1842–1859.e18.

Marcus J. H., C. Posth, H. Ringbauer, L. Lai, R. Skeates, *et al.*, 2020 Genetic history from the Middle Neolithic to present on the Mediterranean island of Sardinia. Nat. Commun. 11: 939.

Maróti Z., E. Neparáczki, O. Schütz, K. Maár, G. I. B. Varga, *et al.*, 2022 The genetic origin of Huns, Avars, and conquering Hungarians. Curr. Biol. 32: 2858–2870.e7.

Martin S. H., J. W. Davey, and C. D. Jiggins, 2015 Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol. Biol. Evol. 32: 244–257.

Mathieson I., S. Alpaslan-Roodenberg, C. Posth, A. Szécsényi-Nagy, N. Rohland, *et al.*, 2018 The genomic history of southeastern Europe. Nature 555: 197–203.

McVean G., 2009 A genealogical interpretation of principal components analysis. PLoS Genet. 5: e1000686.

Mitchell K. J., and N. J. Rawlence, 2021 Examining Natural History through the Lens of Palaeogenomics. Trends Ecol. Evol. 36: 258–267.

Moots H. M., M. Antonio, S. Sawyer, J. P. Spence, V. Oberreiter, *et al.*, 2023 A genetic history of continuity and mobility in the Iron Age central Mediterranean. Nat Ecol Evol 7: 1515–1524.

Narasimhan V. M., N. Patterson, P. Moorjani, N. Rohland, R. Bernardos, *et al.*, 2019 The formation of human populations in South and Central Asia. Science.

Nelson D., J. Kelleher, A. P. Ragsdale, C. Moreau, G. McVean, *et al.*, 2020 Accounting for long-range correlations in genome-wide simulations of large cohorts. PLoS Genet. 16: e1008619.

Nielsen S. V., A. H. Vaughn, K. Leppälä, M. J. Landis, T. Mailund, *et al.*, 2023 Bayesian inference of admixture graphs on Native American and Arctic populations. PLoS Genet. 19: e1010410.

Ning C., T. Li, K. Wang, F. Zhang, T. Li, *et al.*, 2020 Ancient genomes from northern China suggest links between subsistence changes and human migration. Nat. Commun. 11: 2700.

Olalde I., S. Mallick, N. Patterson, N. Rohland, V. Villalba-Mouco, *et al.*, 2019 The genomic history of the Iberian Peninsula over the past 8000 years. Science 363: 1230–1234.

Oliveira S., K. Nägele, S. Carlhoff, I. Pugach, T. Koesbardiati, *et al.*, 2022 Ancient genomes from the last three millennia support multiple human dispersals into Wallacea. Nat Ecol Evol 6: 1024–1034.

Papac L., M. Ernée, M. Dobeš, M. Langová, A. B. Rohrlach, *et al.*, 2021 Dynamic changes in genomic and social structures in third millennium BCE central Europe. Sci Adv 7.

Patterson N., A. L. Price, and D. Reich, 2006 Population Structure and Eigenanalysis. PLoS Genet. 2: e190.

Patterson N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, *et al.*, 2012 Ancient admixture in human history. Genetics 192: 1065–1093.

Patterson N., M. Isakov, T. Booth, L. Büster, C.-E. Fischer, *et al.*, 2022 Large-scale migration into Britain during the Middle to Late Bronze Age. Nature 601: 588–594.

Peter B. M., 2016 Admixture, Population Structure, and F-Statistics. Genetics 202: 1485–1501.

Peter B. M., 2022 A geometric relationship of F2, F3 and F4-statistics with principal component analysis. PHILOSOPHICAL TRANSACTIONS B.

Pickrell J. K., and J. K. Pritchard, 2012 Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8: e1002967.

Reich D., A. L. Price, and N. Patterson, 2008 Principal component analysis of genetic data. Nat. Genet. 40: 491–492.

Reich D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh, 2009 Reconstructing Indian population history. Nature 461: 489–494.

Reich D., N. Patterson, D. Campbell, A. Tandon, S. Mazieres, *et al.*, 2012 Reconstructing Native American population history. Nature 488: 370–374.

Salazar L., R. Burger, J. Forst, R. Barquera, J. Nesbitt, *et al.*, 2023 Insights into the genetic histories and lifeways of Machu Picchu's occupants. Sci Adv 9: eadg3377.

Schmid C., and S. Schiffels, 2023 Estimating human mobility in Holocene Western Eurasia with large-scale ancient genomic data. Proc. Natl. Acad. Sci. U. S. A. 120: e2218375120.

Schuenemann V. J., A. Peltzer, B. Welte, W. P. van Pelt, M. Molak, *et al.*, 2017 Ancient Egyptian mummy

genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. Nat. Commun. 8: 15694.

Sirak K. A., D. M. Fernandes, M. Lipson, S. Mallick, M. Mah, *et al.*, 2021 Social stratification without genetic differentiation at the site of Kulubnarti in Christian Period Nubia. Nat. Commun. 12: 7283.

Skoglund P., J. C. Thompson, M. E. Prendergast, A. Mittnik, K. Sirak, *et al.*, 2017 Reconstructing Prehistoric African Population Structure. Cell 171: 59–71.e21.

Skourtanioti E., Y. S. Erdal, M. Frangipane, F. Balossi Restelli, K. A. Yener, *et al.*, 2020 Genomic History of Neolithic to Bronze Age Anatolia, Northern Levant, and Southern Caucasus. Cell 181: 1158–1175.e28.

Skourtanioti E., H. Ringbauer, G. A. Gnecchi Ruscone, R. A. Bianco, M. Burri, *et al.*, 2023 Ancient DNA reveals admixture history and endogamy in the prehistoric Aegean. Nat Ecol Evol 7: 290–303.

Slatkin M., and F. Racimo, 2016 Ancient DNA and human history. Proc. Natl. Acad. Sci. U. S. A. 113: 6380–6387.

Soraggi S., and C. Wiuf, 2019 General theory for stochastic admixture graphs and F-statistics. Theor. Popul. Biol. 125: 56–66.

Spyrou M. A., K. I. Bos, A. Herbig, and J. Krause, 2019 Ancient pathogen genomics as an emerging tool for infectious disease research. Nat. Rev. Genet. 20: 323–340.

Taylor W. T. T., P. Librado, C. J. American Horse, C. Shield Chief Gover, J. Arterberry, *et al.*, 2023 Early dispersal of domestic horses into the Great Plains and northern Rockies. Science 379: 1316–1323.

Tricou T., E. Tannier, and D. M. de Vienne, 2022 Ghost Lineages Highly Influence the Interpretation of Introgression Tests. Syst. Biol. 71: 1147–1158.

Wang C.-C., S. Reinhold, A. Kalmykov, A. Wissgott, G. Brandt, *et al.*, 2019 Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. Nat. Commun. 10: 590.

Wang K., I. Mathieson, J. O'Connell, and S. Schiffels, 2020 Tracking human population structure through time from whole genome sequences. PLoS Genet. 16: e1008552.

Wang C.-C., H.-Y. Yeh, A. N. Popov, H.-Q. Zhang, H. Matsumura, *et al.*, 2021 Genomic insights into the formation of human populations in East Asia. Nature 591: 413–419.

Wibowo M. C., Z. Yang, M. Borry, A. Hübner, K. D. Huang, *et al.*, 2021 Reconstruction of ancient microbial genomes from the human gut. Nature 594: 234–239.

Williams M., and J. Teixeira, 2020 A genetic perspective on human origins. Biochem. 42: 6–10.

Wood S. N., 2004 Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. J. Am. Stat. Assoc. 99: 673–686.

Yang M. A., X. Fan, B. Sun, C. Chen, J. Lang, *et al.*, 2020 Ancient DNA indicates human population shifts and admixture in northern and southern China. Science 369: 282–288.

Yüncü E., U. Işıldak, M. P. Williams, C. D. Huber, O. Flegontova, *et al.*, 2023 False discovery rates of qpAdm-based screens for genetic admixture. bioRxiv 2023.04.25.538339.

Zheng Y., and A. Janke, 2018 Gene flow analysis method, the D-statistic, is robust in a wide parameter space. BMC Bioinformatics 19: 10.